VOLUME 103 PART B NUMBER 7 JANUARY 1956

The Proceedings OF THE INSTITUTION OF ELECTRICAL ENGINEERS

FOUNDED 1871: INCORPORATED BY ROYAL CHARTER 1921

PART B

RADIO AND ELECTRONIC ENGINEERING
(INCLUDING COMMUNICATION ENGINEERING)

SAVOY PLACE . LONDON W.C.2

Price Seven Shillings and Sixpence

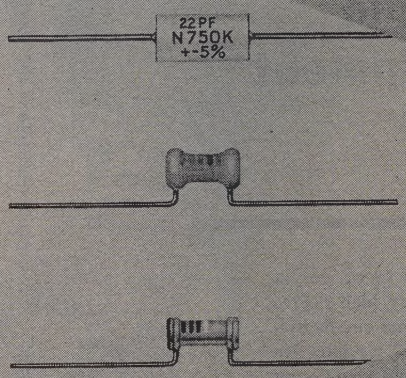
The Institution of Electrical Engineers

ERIE* Tubular Ceramicons*

STYLES K-M TROPICALISED CERAMIC ENCASED

STYLES AD-CD TROPICALISED INSULATED COATING

STYLES A-C TROPICALISED NON-INSULATED COATING



Illustrations actual size

14 CLOSELY DEFINED TEMPERATURE COEFFICIENTS

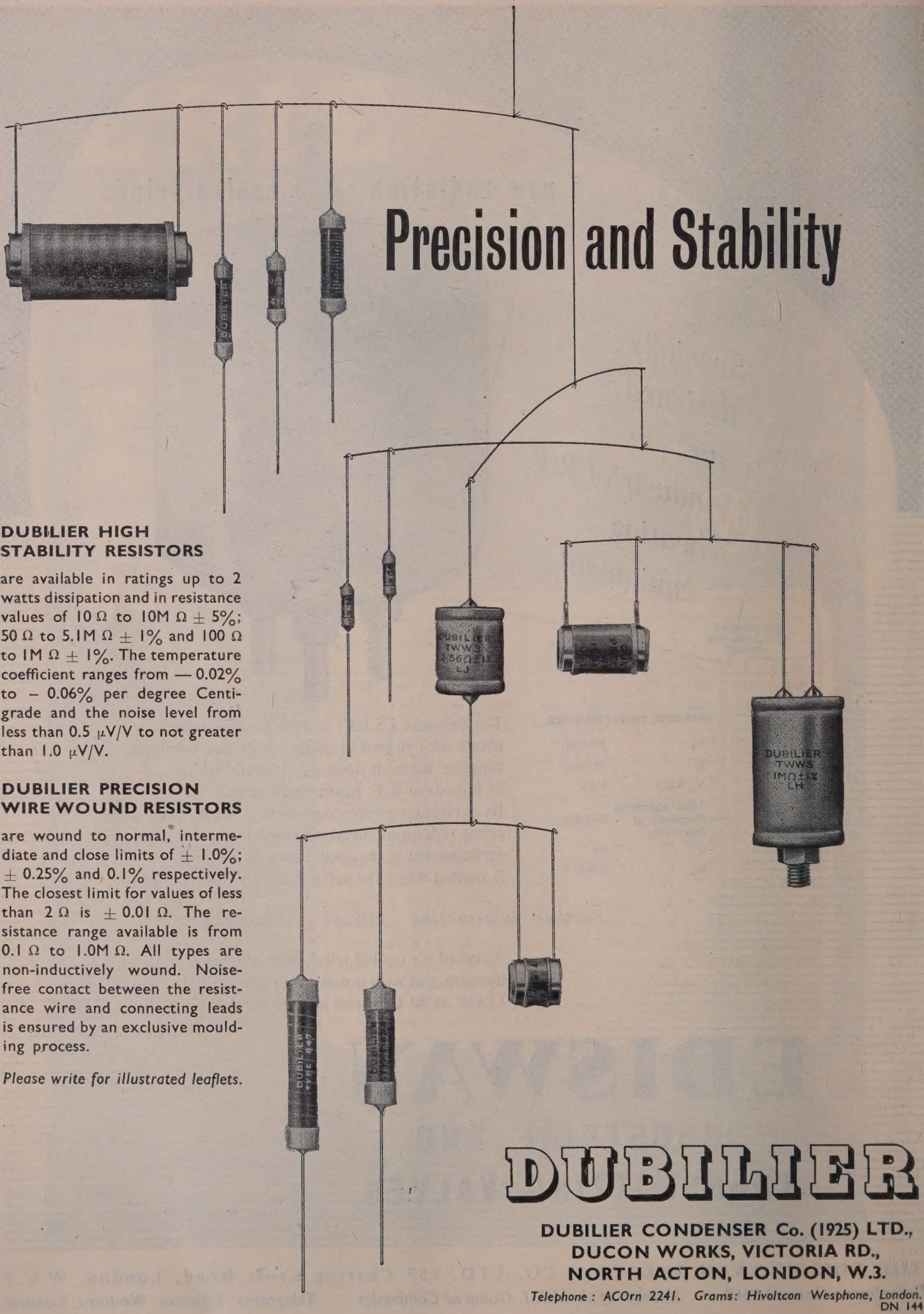
EVR*

STYLE	TROPICALISED CERAMIC ENCASED			TROPICALISED INSULATED COATING			TROPICALISED NON-INSULATED COATING		
	K	L	M	AD	BD	CD	A	B	C
DIM. L MAX.	0.540"	0.800"	1.300"	0.460"	0.710"	1.250"	0.385"	0.650"	1.120"
DIM. D MAX.	0.255"	0.255"	0.375"	0.240"	0.240"	0.315"	0.200"	0.200"	0.250"
CAPACITANCE RANGE PF									
P100	Up to 10	11 – 22	23 – 90	Up to 10	11 – 22	23 – 90	Up to 10	11 – 22	23 – 90
NPO	Up to 18	19 – 42	43 – 130	Up to 18	19 – 42	43 – 130	Up to 18	19 – 42	43 – 130
N030	Up to 18	19 – 45	46 – 133	Up to 18	19 – 45	46 – 133	Up to 18	19 – 45	46 – 133
N080	Up to 21	22 – 53	54 – 155	Up to 21	22 – 53	54 – 155	Up to 21	22 – 53	54 – 155
N150	Up to 24	25 – 60	61 – 176	Up to 24	25 – 60	61 – 176	Up to 24	25 – 60	61 – 176
N220	Up to 27	28 – 66	67 – 194	Up to 27	28 – 66	67 – 194	Up to 27	28 – 66	67 – 194
N330	Up to 30	31 – 73	74 – 215	Up to 30	31 – 73	74 – 215	Up to 30	31 – 73	74 – 215
N470	Up to 36	37 – 88	89 – 260	Up to 36	37 – 88	89 – 260	Up to 36	37 – 88	89 – 260
N750	2 – 62	63 – 120	121 – 460	2 – 62	63 – 120	121 – 460	2 – 62	63 – 120	121 – 460
N1500	10 – 60	61 – 147	148 – 430	10 – 60	61 – 147	148 – 430	10 – 60	61 – 147	148 – 430
N2200	20 – 95	96 – 235	236 – 688	20 – 95	96 – 235	236 – 688	20 – 95	96 – 235	236 – 688
N3300	40 – 149	150 – 368	369 – 1080	40 – 149	150 – 368	369 – 1080	40 – 149	150 – 368	369 – 1080
N4700	60 – 328	329 – 809	810 – 2370	60 – 328	329 – 809	810 – 2370	60 – 328	329 – 809	810 – 2370
N5600	80 – 478	479 – 1180	1181 – 3440	80 – 478	479 – 1180	1181 – 3440	80 – 478	479 – 1180	1181 – 3440

ERIE* Resistor Ltd

* Registered Trade Marks

Carlisle Road, The Hyde, London, N.W.9., England. Telephone: COLindale 8011.
Factories: London and Great Yarmouth, England: Trenton, Ontario, Canada: Erie, Pa., and Holly Springs, Miss., U.S.A.



Precision and Stability

DUBILIER HIGH STABILITY RESISTORS

are available in ratings up to 2 watts dissipation and in resistance values of 10 Ω to 10M Ω ± 5%; 50 Ω to 5.1M Ω ± 1% and 100 Ω to 1M Ω ± 1%. The temperature coefficient ranges from — 0.02% to – 0.06% per degree Centigrade and the noise level from less than 0.5 μV/V to not greater than 1.0 μV/V.

DUBILIER PRECISION WIRE WOUND RESISTORS

are wound to normal, intermediate and close limits of ± 1.0%; ± 0.25% and 0.1% respectively. The closest limit for values of less than 2 Ω is ± 0.01 Ω. The resistance range available is from 0.1 Ω to 1.0M Ω. All types are non-inductively wound. Noise-free contact between the resistance wire and connecting leads is ensured by an exclusive moulding process.

Please write for illustrated leaflets.

DUBILIER

DUBILIER CONDENSER Co. (1925) LTD.,
DUCON WORKS, VICTORIA RD.,
NORTH ACTON, LONDON, W.3.

Telephone : ACOrn 2241. Grams: Hivoltcon Wesphone, London.

DN 144

G.E.C.

Britain's microwave leaders announce

The New 60/120-CIRCUIT U.H.F. TELEPHONY SYSTEM SPO 5500

THIS frequency-modulated system, conveying either 60 or 120 circuits, operates in the 1700–2300 Mc/s band. Long systems show a minimum of modulation distortion since non-demodulating repeater stations are used. The system handles two super groups of 60 circuits each, with a total signal/noise performance in the worst channel only 6db below that recommended for C.C.I.F. international coaxial cable networks.

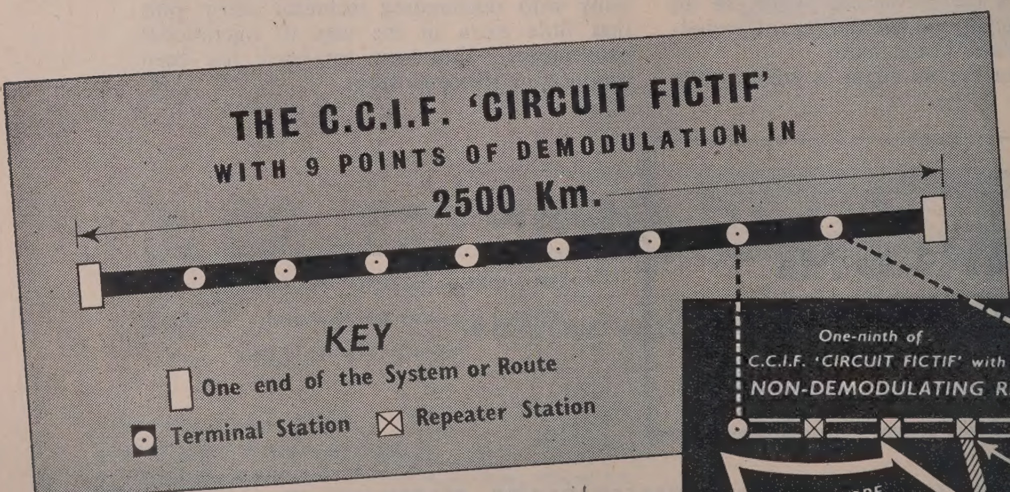
The most modern construction practice permits all panels to slide into place on guides, being connected into service by plug-in sockets. Wiring is not disconnected for the removal of a panel.

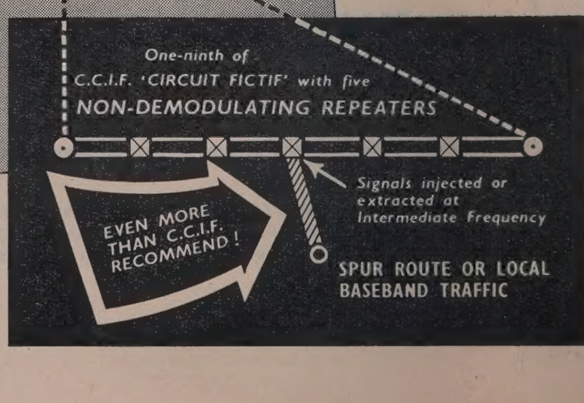
Each rack has a meter panel, giving readings of anode and grid currents, crystal currents, R F amplifier output power, and all non-mains voltages.

No voltage higher than 300V is encountered in the equipment.

The use of coaxial cable for feeders eliminates the expense of wave-guides.

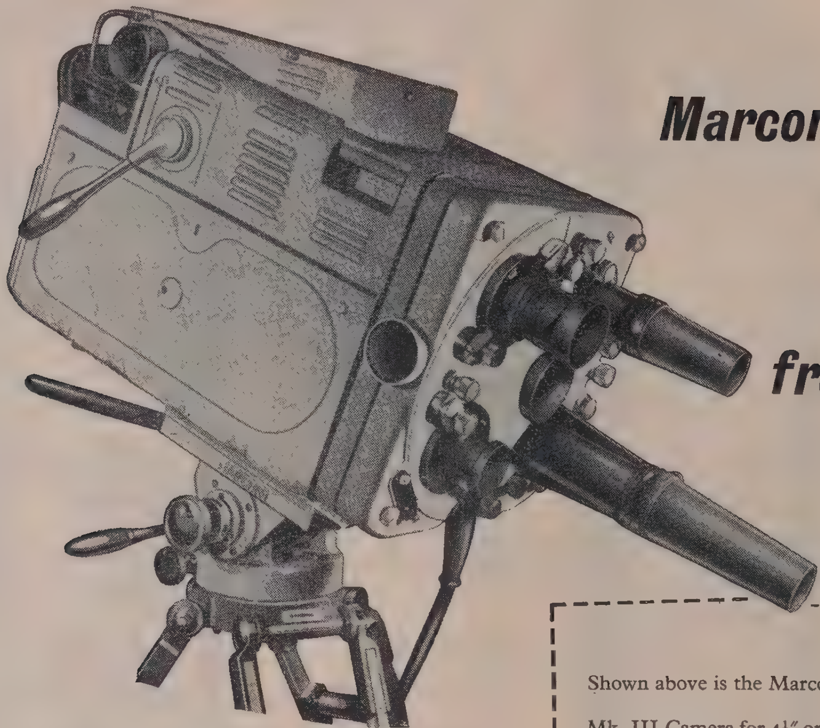
A rack complete with transmitter and receiver is single-sided, so that two racks may be mounted side-by-side or back-to-back. They are economical of floor space, occupying only 20½ in. × 8½ in.





● *For further details write for Standard Specification SPO. 5500.*

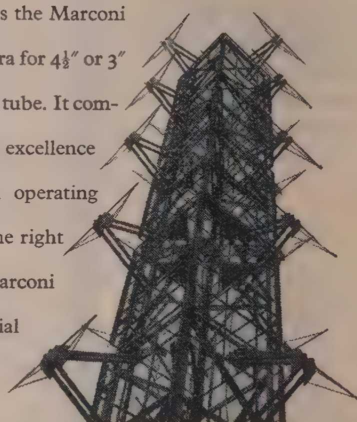
The cost of the radio equipment is about half that of copper wire alone to provide 60 circuits by 12-circuit carrier systems on open-wire routes.



Marconi Complete TELEVISION SYSTEMS from Camera to Aerial

Both for studio and outside broadcasting of television programmes Marconi equipment is required at every stage of production and transmission. Cameras, Picture and Waveform Monitors, Vision Mixers, Telecine Equipment where film sequences are included, Microwave Links, Transmitters and Aerial must all be matched as components of a completely integrated system with a designed performance. The long experience and advanced technical knowledge of the Company's Broadcasting Division are at the disposal of all who are responsible for television, throughout the world.

Shown above is the Marconi Mk. III Camera for 4½″ or 3″ image orthicon tube. It combines technical excellence with optimum operating facilities. On the right is a typical Marconi Television Aerial array.



Marconi Television Equipment is installed in every one of the B.B.C.'s Television stations and has been supplied to countries in North and South America, Europe and Asia. Compatible colour television was first demonstrated in Britain by Marconi's.

Lifeline of communication

MARCONI

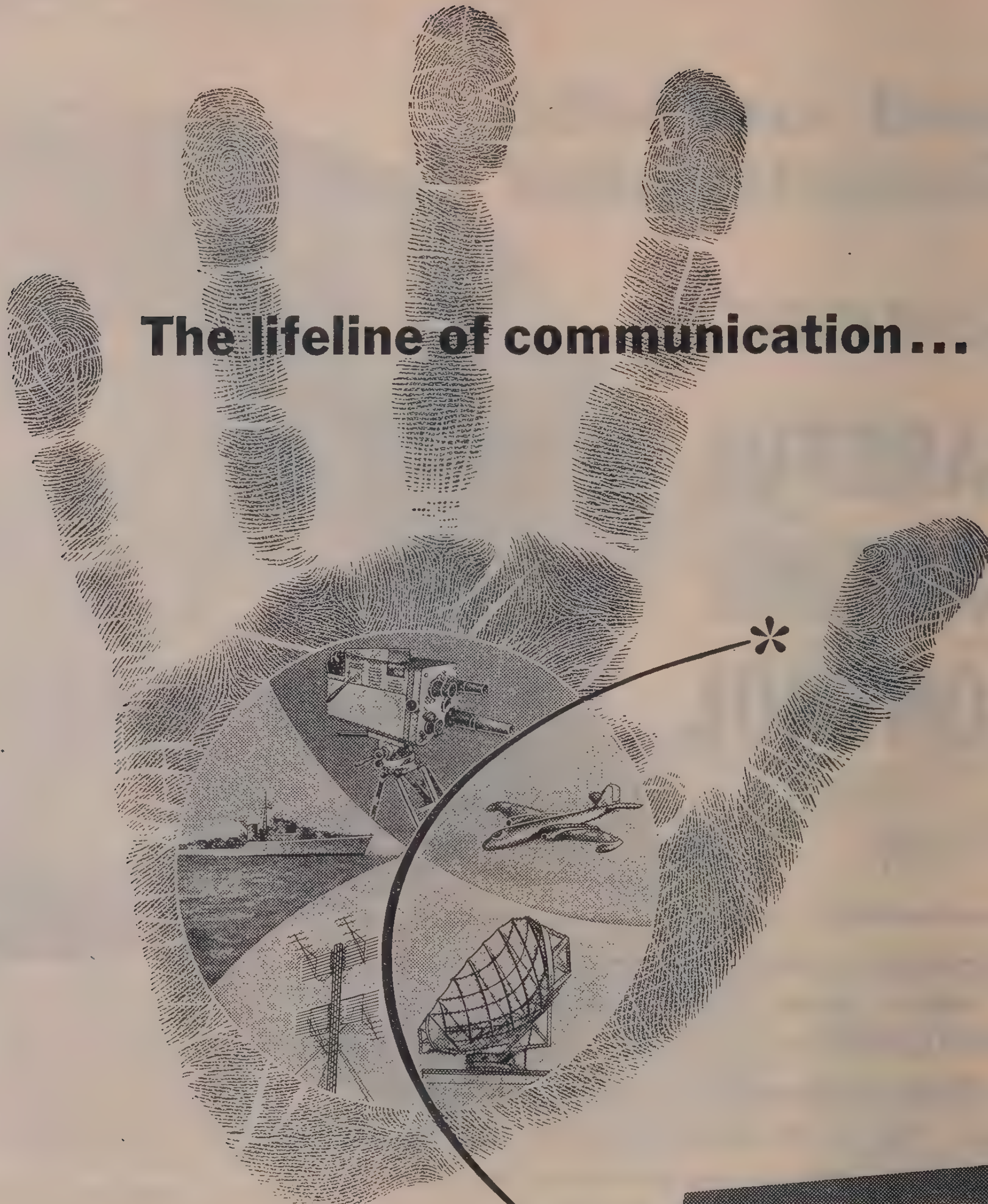
Complete Broadcasting and Television Systems

MARCONI'S WIRELESS TELEGRAPH CO. LTD., CHELMSFORD, ESSEX

Partners in progress with the 'ENGLISH ELECTRIC' *Company Limited*

LB 5

The lifeline of communication...



• More than forty civil airlines and twenty air forces fit Marconi air radio equipment. Airports all over the world rely on Marconi ground installations • The services have entrusted integrated radar air and surface defence networks both at home and overseas to Marconi's • 75% of the countries in the world operate Marconi Broadcasting and Television equipment. • 80 countries have Marconi equipped telegraph and communications systems. • All the radio approach and marker beacons round the coasts of Britain have been supplied by Marconi's.

SYSTEM PLANNERS,
ELECTRONIC ENGINEERS,
DESIGNERS AND MANUFACTURERS
OF AERONAUTICAL, BROADCASTING
COMMUNICATION AND MARITIME
RADIO EQUIPMENT
TELEVISION EQUIPMENT
RADAR AND NAVIGATIONAL AIDS

MARCONI

on land, at sea and in the air

MARCONI'S WIRELESS TELEGRAPH COMPANY LIMITED · CHELMSFORD · ESSEX

LG 6


CATHODEON
Quartz Crystals
FOR RELIABLE FREQUENCY CONTROL
Frequency Range 2,000—20,000 kc/s
Our range now includes crystals for close tolerance requirements
Enquiries are invited for overtones up to 60 Mc/s
CATHODEON CRYSTALS LIMITED
LINTON · CAMBRIDGESHIRE · Telephone LINTON 223

THE INSTRUMENT MODEL

Specially designed for soldering operations in the compact assemblies used in present day radio, television and electronic industries.

Weight $3\frac{1}{4}$ oz. (excluding flexible).

Length 9 in. 25 Watts.

Voltage Range
12, 24, 50, 100/110, 120/130, 200/220 and 220/240.

Interesting features

1. Bit $\frac{3}{16}''$ diameter, simple to replace.
2. Steel cased element, also replaceable.
3. Detachable hook for suspending iron when not in use.
4. Moulded two part handle, remains cool in use.
5. Six ft. Henley Flexible.

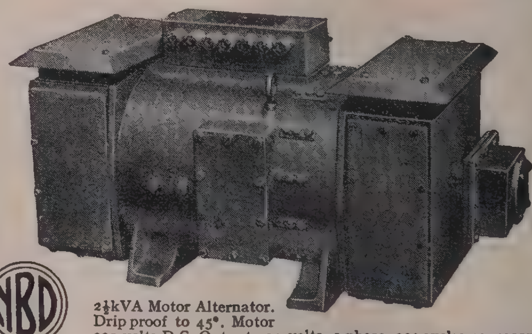
W. T. HENLEY'S TELEGRAPH WORKS CO., LTD.,
51/53, Hatton Garden, London, E.C.1

NEWTON-DERBY
ELECTRICAL EQUIPMENT

High Frequency Alternators

(Send for Publication No. 1003/2)

Also makers of Rotary Transformers and Converters, Wind and Engine-Driven Aircraft Generators, High Tension D.C. Generators, and Automatic Carbon Pile Voltage Regulators.



NBD

$2\frac{1}{2}$kVA Motor Alternator. Drip proof to 45°. Motor 220 volts D.C. Output 120 volts. 3 phase. 333 cycles per second Motor includes an automatic constant speed governor. Weight 450 lb.

NEWTON BROTHERS (DERBY) LTD

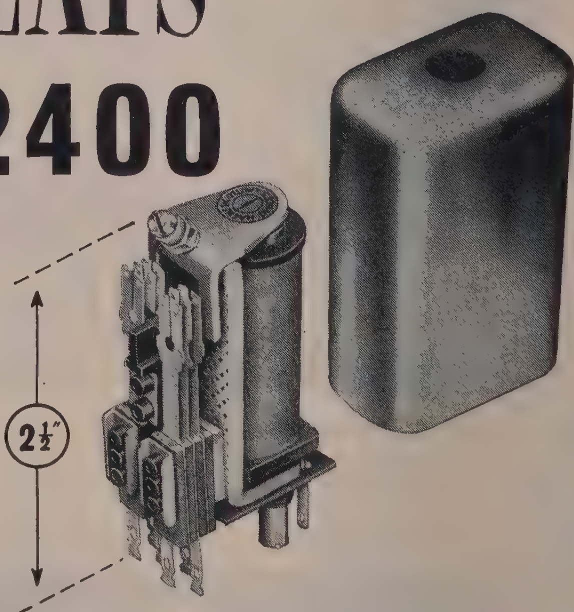
HEAD OFFICE & WORKS: ALFRETON ROAD, DERBY
TELEPHONE: DERBY 47676 (4 lines) TELEGRAMS: DYNAMO, DERBY
LONDON OFFICE: IMPERIAL BUILDINGS, 56 KINGSWAY W.C.2

RELAYS
SERIES 2400

A Relay of noteworthy dimensions, designed in size and performance to suit present-day electronic equipment. The new 2400 Relay is available with twin light duty or single heavy duty contacts. When fitted with a 10,000 ohm coil, the pull-in is approximately 4 milliamperes; contact pressure and clearance have not been sacrificed to achieve this sensitivity.

DIMENSIONS: Above chassis $2\frac{1}{2}''$ high x $1''$ wide x $1\frac{5}{8}''$ deep.

WEIGHT: $4\frac{1}{2}$ ounces.



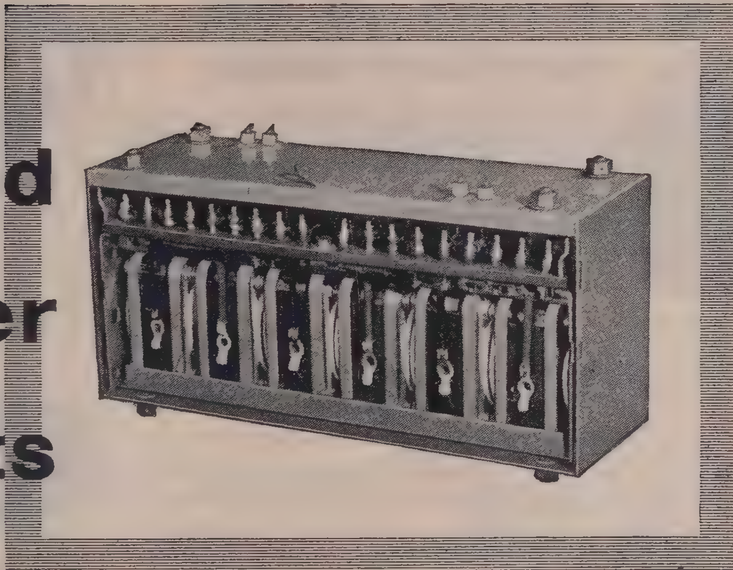
Telephone: Newmarket 3181-2-3 Telegrams: Magnetic, Newmarket

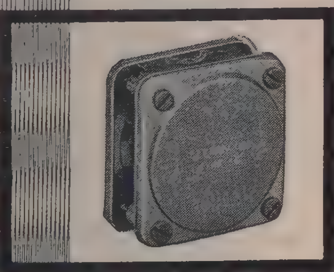
MAGNETIC DEVICES LIMITED

MAGNETIC DEVICES LTD
EXNING ROAD, NEWMARKET

MD 3A/2

Improved filter units





with Ferroxcube pot cores

1 *High performance combined with small size and light weight.*

2 *Designed and built to customers' individual requirements.*

3 *Long term stability, even under conditions of temperature variation.*

High quality electrical filter units built around Ferroxcube cores can now be supplied to communications equipment designers' individual specifications. These filter units have significant advantages over comparable types designed without the use of Ferroxcube, particularly in the frequency range 300 c/s to 500 kc/s. For audio frequencies the use of Ferroxcube cores permits the winding of compact coils with very high inductances. This results in a considerable reduction in the size and cost of the associated condensers and hence of the filter unit as a whole. The high Q values obtained for a given volume, especially above 10 kc/s, enable sharp cut off characteristics and low pass-band losses to be achieved, while negligible stray flux facilitates the production of compact and mechanically robust filters. Electrical filter units are among a number of high quality components now being made available by Mullard. Full details of the complete series of components will be gladly supplied upon request.

Mullard

'Ticonal' alloy permanent magnets
Magnadur permanent ceramic magnets
Ferroxcube ferro-magnetic cores.

MULLARD LTD · COMPONENTS DIVISION · CENTURY HOUSE · SHAFTESBURY AVENUE · WC2

(M1463A)

The Benevolent Fund

HAVE YOU YET RESPONDED TO THE APPEAL FOR CONTRIBUTIONS TO THE

HOMES FUND

THE COURT OF GOVERNORS HOPE THAT EVERY MEMBER WILL CONTRIBUTI TO THIS WORTHY OBJECT

Contributions may be sent by post to

THE INCORPORATED BENEVOLENT FUND OF THE INSTITUTION OF ELECTRICAL ENGINEERS
SAVOY PLACE LONDON W.C.2

or may be handed to your Local Hon. Treasurer.

LOCAL HON. TREASURERS OF THE FUND

East Midland Centre	*R. C. Woods*
Irish Branch	*A. Harkin, M.E.*
Mersey and North Wales Centre	*D. A. Picken*
North-Eastern Centre	*D. R. Parsons*
North Midland Centre	*J. G. Craven*
Sheffield Sub-Centre	*F. Seddon*
North-Western Centre	*W. E. Swale*
North Lancashire Sub-Centre	*G. K. Alston, B.Sc.(Eng.)*
Northern Ireland Centre	*G. H. Moir, J.P.*
Scottish Centre	*R. H. Dean, B.Sc.Tech.*
North Scotland Sub-Centre	*P. Philip*
South Midland Centre	*W. E. Clark*
Rugby Sub-Centre	*H. Orchard*
Southern Centre	*G. D. Arden*
Western Centre (Bristol)	*A. H. McQueen*
Western Centre (Cardiff)	*D. J. Thomas*
South-Western Sub-Centre	*W. E. Johnson*
West Wales (Swansea) Sub-Centre	*O. J. Mayo*

THE PROCEEDINGS OF
THE INSTITUTION OF ELECTRICAL ENGINEERS

EDITED UNDER THE SUPERINTENDENCE OF W. K. BRASHER, C.B.E., M.A., M.I.E.E., SECRETARY

VOL. 103. PART B. No. 7. JANUARY 1956

Paper No. 1953
Oct. 1955

INAUGURAL ADDRESS

By Sir GEORGE H. NELSON, Bt., M.I.Mech.E., President

(*Address delivered before* THE INSTITUTION *6th October*, 1955.)

I am greatly touched by the honour conferred on me by the ouncil and members of The Institution in electing me President r the ensuing year, and thank you all for this great compliment ou pay me. I shall, with all your help, do everything possible uphold the traditions and high standards of this great stitution.

That it should happen in this particular year is especially easing to me, coming as it does on the 50th anniversary of y admission to The Institution as a Student Member.

The theme of my Address will be the great prospects and the sponsibilities of electrical engineers in contributing in the ture to the benefit of man everywhere. The subject falls into ree sections—past, present and future.

I will start the picture of the past with our first President, harles William Siemens, who, it is most interesting to recall, ppened to have founded a section of my own Company. I d the pleasure, some two years ago, of presenting to The stitution facsimiles of his letters giving fascinating pictures the technical and social life of his time. From this correspondence letter No. 43 shows that, as long ago as 1873, rliament had moved for the appointment of a Committee "to quire into the causes of present dearness and scarcity" of al!

His Address to the first meeting of our Institution—then lled The Society of Telegraph Engineers—in February, 1872 ne Society having actually been founded in 1871), was most ophetic. He pointed out that, what to many then seemed to separate compartments of science and engineering, were ally one, involving a vast range of problems in theoretical ysics, applied chemistry, engineering and industrial management, developing into the whole field of electrical engineering we know it to-day. He said:

> There is hardly a problem in electrical science that is not of practical interest to the telegraph engineer . . . The phenomena of electrification and polarization, of specific induction and conduction, the laws regulating the electrical wave, the influences of . . . temperature on conduction . . . the potential force residing in a coil of wire of a given form, when traversed by a current, involve questions belonging just as much to pure physical science as to the daily practice of the telegraph engineer . . .

d went on to refer to

> . . . questions of selection of materials for conduction or insulation . . . and apparatus for producing and directing . . . electrical current which . . . call into play considerations . . . of purely mechanical import. . . . I would go further and include statistical information.

He concluded:

> These remarks may suffice to show how great is the field of our activity and how much remains to be accomplished notwithstanding the extraordinary progress of which we are apt to boast.

After this address Mr. C. F. Varley, F.R.S., made most interesting observations, as follows:

> Mr. President . . .
> After your Address, Sir, no one will fail to see . . . this Society . . . will gradually, by natural selection, develop more into an electrical society . . . the moment it is understood that all papers on electricity, or bearing directly upon the development of electrical science are admitted . . . because it will be found ultimately to embrace every operation in nature.

Our Institution, starting in this atmosphere of the widening orbit of electrical science and engineering was to see its original name changed twice in less than eight years, first to "The Society of Telegraph Engineers and Electricians," in 1880, and then to "The Institution of Electrical Engineers," in 1888, when the professional term and title of Electrical Engineer was established.

The work of The Institution and of our profession had progressed so much by 1921 that a Royal Charter was applied for—and was granted, a very fitting honour and recognition of The Institution's central role in establishing high professional standards and propagating and encouraging the science and technology of electricity which has wrought such a revolution in the lives of people everywhere.

Among The Institution's many functions its publication of *Science Abstracts* must rank high in importance, and it is interesting that, at the first meeting in 1872, Prof. Foster, F.R.S., drew attention to the need for a service of this kind.

The three great principles of The Institution's work and policy laid down are:

(i) To act as a learned society and to provide the means to exchange knowledge in electrical matters.
(ii) To act as a qualifying body in fixing the standards of knowledge of electrical engineering in its theory and practice.
(iii) To determine guidance of ethical conduct in our profession.

Our predecessors have striven from the very beginning for the establishment of standards at the highest level, and we and our successors will, I am sure, jealously guard and uphold them.

To refresh your memories on the progress that has been made

in the generation, distribution and use of electrical energy, I would mention that in 1905 many generating plants were of the order of only a few hundred kilowatts, powered by reciprocating prime movers, and that the total generating capacity in Great Britain was 1 700 MW, being only $\frac{1}{20}$ kW per head of population. At the same time in the United States the total generating capacity was 6 800 MW—only a little over $\frac{1}{12}$ kW per head of population.

What immense advances over these figures are shown by to-day's statistics, for by June of this year the total generating capacity in Great Britain had reached 25 500 MW, or $\frac{1}{2}$ kW per head of population—a tenfold increase.

The capacities of individual generating sets have risen in 50 years from a few hundred kilowatts to over 100 MW and, quite recently, sets of 200 MW have been ordered.

In the United States the installed capacity, in the same time, has risen to over 121 000 MW or $\frac{3}{4}$ kW per head of population, and this includes sets larger than 200 MW.

These developments have brought about a steady fall in the cost of electric power in comparison with the cost of living over the past 30 years, amounting to more than 45% in this country and 50% in the United States. The advances in the design of electrical equipment have come from great imagination, courage and enormous expenditure on research and development by the heavy electrical manufacturing industry financed almost entirely from its own resources.

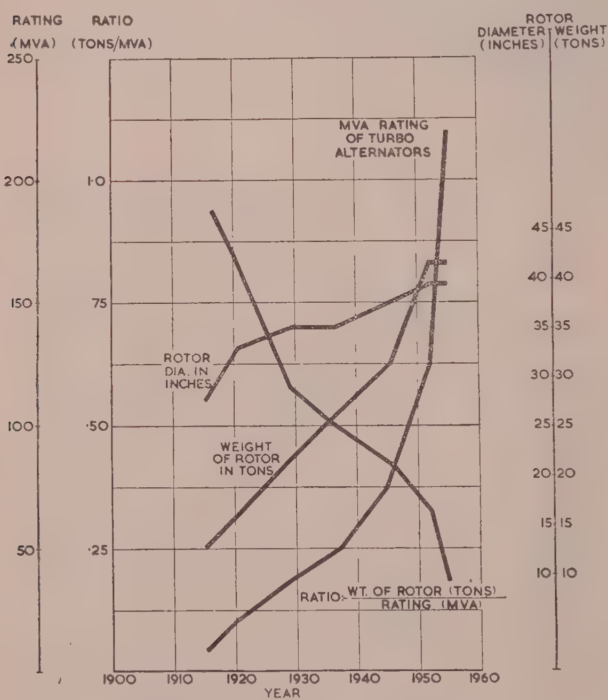


Fig. 1.—Development of turbo-alternators.

Fig. 1 indicates some of the technical changes that have taken place in turbo-alternators in the past forty years. It will be seen that although the output of machines has increased more than 20 times, the rotor weight has gone down from nearly 1·0 ton/MVA to 0·2 ton/MVA, i.e. it has been reduced by 80%. Both these changes have been made possible by new techniques—including the introduction of hydrogen cooling—arising from research. Similar changes have taken place with the steam turbine, through the use of higher steam pressures and temperatures.

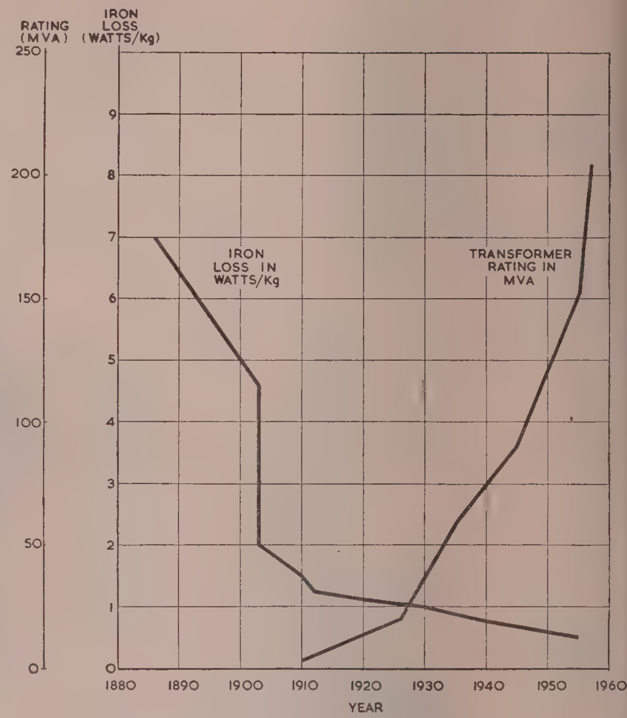


Fig. 2.—Development of transformers.

In distribution, similar spectacular advances have been made. Fig. 2 indicates some of the changes in the design of transformers. The curve shows what a big advance took place about 1905 with the introduction of silicon steel—invented by Hadfield. More improvement is now obtained through grain orientation during the manufacture of the silicon iron. The result has been that the power losses per pound of core iron have been reduced to one-quarter of what they were 50 years ago, and ratings have gone up from 5 MVA to more than 200 MVA, with a rise in transmission voltages from about 6 kV to 400 kV.

As a result of continuous and intensive research and development, corresponding progress has been made with switchgear for dealing with equally spectacular increases in transmission voltages from 6 kV up to 400 kV. Similarly interruption ratings have risen from 25 MVA to 10 000 MVA.

In the telecommunication field progress has been staggering. Fifty years ago few homes had a telephone, Marconi's monumental achievement in transatlantic radiocommunication was only four years old and there was no system of radiocommunication as we know it to-day. The first patent for wireless was only taken out in 1896, and a British Company, formed the following year, was the first to bring wireless telegraphy to the world.

The first public broadcast of news and entertainment did not come until 1920, and the first television service until 1936. Colour television could come at any time now, when the economic position allows it, since the technical problems are mainly solved.

Progress in the application of electrical and electronic control to many important industries is increasing production in a spectacular manner, and also improving quality. An outstanding example is the rolling of sheet steel which, up to 30 years ago was produced from steam-driven mills in small sheets at a rate of about 2·5 m.p.h. To-day, sheet steel of much more accurate dimensions is produced in coil lengths of 15 miles, leaving the rolls at a rate of nearly 60 m.p.h. The control of liquid and gas

flow and the control of temperatures in the delicate processes of oil refining is another field of important application. What further advance in productivity in other industries will result from a combination of electric power and electronic methods of process and machine control one can only guess, but the fields for such application are unlimited.

Next comes the gain in users' time and in efficiency, by the use of the electronic computing machine, in the solution of research and design problems. These machines are already saving man-weeks and man-months of engineers' and mathematicians' time, solving problems in research and design in electrical, aerodynamic and many other branches of science and engineering. In fact, the availability of these machines has enabled problems to be solved which, up to the present, had not been found possible, owing to the effort needed for their solution.

With these brief comments on the achievements of the past and present, I will now turn to the prospects of the future together with the problems that will be associated with that progress.

From my extensive travels oversea, and from discussions with those carrying the highest responsibility in the many countries visited, particularly since the end of the 1939–45 War, I have found an almost unanimous determination—not just a hope—I repeat, a determination, of all nations—whatever their race, creed or state of development—to improve their standard of living, and they all realize that the key to this lies in the greater application of electric power to the source of their national economy, including agriculture, communications and industry. In other words, economically, they all have an expansionist policy.

To the electrical engineering profession and industry this brings increased responsibility now, with a steadily increasing demand for engineers and for electrical equipment by the pressure of more and more people everywhere, as the population of the world increases.

Let us briefly review the prospects of the growth of the world population. Estimates given by Palmer Putnam in his book,* and also statistics issued by the United Nations, show that the population of the world has risen from 1 100 million in 1850 to 2 400 million in 1950. Over half of that population are living in areas where there is an immediate and continuing pressure for improvement in economic standards.

It is estimated that by the year 2050 the world's population will have increased to $2\frac{1}{2}$ times the 1950 figure, reaching the gigantic total of 6000 million.

These estimates of world population cannot, of course, be taken too factually, because of the difficulty of obtaining reliable censuses in the under-developed countries; they can be nothing more than rough estimates, but they suffice to indicate that both living standards and world population, steadily and substantially rising, will increase the demand for more and more electric power.

At the recent Geneva Atomic Energy Conference, Prof. Robinson and Dr. Daniel said that by the year 2000 the world consumption of energy would be not less than the equivalent energy of 7 500 million tons of coal a year, representing a threefold increase in total energy consumption in the next 50 years.

These prospects compel us to recognize the needs and compel us to solve the problems which will arise from the steady exhaustion of conventional fuels and basic materials which will be used in meeting the needs of the increased population and the higher standards of living which they will expect.

Fortunately, under the pressure of the need to defend our way of life, we carried on great research and development in the field of nuclear physics through which—by Providence—scientists, technologists and engineers are finding a solution to the exhaustion of conventional fuel in the development of the basis of the means of generation of electric power from nuclear energy.

I had intended to make considerable reference to the bright future in this field, but so much has been said and written, since I framed this Address, that I feel you are all reasonably up to date with the position, and I will simply say, therefore, that the progress made in this field of nuclear energy forms one of those miracles which occur from time to time, at the appropriate moment, to meet the needs of mankind.

The solution of the fuel problem in the generation of electricity is therefore undoubtedly in sight, but we must not for one moment think that this is the only problem with which we are faced in meeting the needs of improving the standards of living of the world and of a growing population. However, just as research and development in nuclear physics provide an alternative source of heat for generating steam, so equal and intensive study, research, and development in fundamental physics will provide alternative sources of supply of other basic materials, and this must be followed up immediately.

The measure of urgency arises from the fact that 90% of the present population of the world has a standard of living below that in the United States and Great Britain, and it has been estimated that if the standard of this 90% were raised overnight to that of Great Britain, the resources of basic materials like iron and copper, etc., would be exhausted in 25 years. Progress, thank goodness, cannot take place at this rate, but it does indicate that the rate of exhaustion may easily become an acute problem in 100 years. One hundred years is a very short period, in normal times, in which to solve problems of such immense and fundamental character as producing appropriate artificial substitutes, and/or producing synthetically basic metals in quantities as the result of recent research into transmutation or synthesis of a heavy element from a lighter one.

Prominent in the field of production of artificial substitutes are the cellulose products. The possibilities in this direction can be measured by realizing that the energy which reaches the earth from the sun every day is estimated to be at least 50 000 times the total energy produced from all man-made engines in the world to-day—i.e. electrical generating plant, motor cars, ships, etc.—and it is this solar energy which provides the base of cellulose pulp.

In visualizing the possibilities of artificial fibres, I would mention that Terylene fibres have physical properties comparable with steel, which indicates that in time means may be found of economically producing materials in suitable forms as a substitute for steel for structural purposes, making it possible to conserve the use of iron ore to make steel for its magnetic or other special properties.

Similarly, a synthetic material has been produced from which is now manufactured, among other things, pipes, for conveying hot, cold or corrosive fluids. This is replacing copper or lead at a fraction of the price and with less liability to corrosion, and so helps to conserve two metals whose supply is rapidly diminishing.

What immense possibilities towards meeting the needs of the future this picture presents, and how important it is that it should also have immediate intensive study and attention.

Although it has no bearing on the value of solar energy which reaches the earth to-day, I cannot conclude my reference to energy from the sun without reminding you that normal fuels, such as coal and oil, used to-day for power generation, received their stored energy from the sun, and, of course, the falls of water generating hydro-electric power spring from water originally evaporated by the sun's energy making use of the earth's contours.

I feel that the Council and the members of The Institution

* "Energy in the Future."

are to be congratulated on their vision in agreeing that The Institution should play its part in forming and becoming a founder of the British Nuclear Energy Conference in association with the Institutions of Civil, Mechanical and Chemical Engineers and the Institute of Physics. To make most rapid progress from research and development in the field of nuclear physics it is necessary to embrace the sciences, the technology and the engineering of all these Institutions, and therefore the establishment of this conference is of great national importance.

It cannot be denied that outstanding benefits have come from development initiated or extended to create the means of destruction. It is now possible that the latest discoveries of our scientists and technicians of the hydrogen bomb may make the risks of war so terrible that man will not face the consequences of the possible annihilation of our species, and therefore some political solution will be substituted for war—anyhow war between industrial nations.

It is indeed providential that science has put such an instrument at the disposal of our statesmen, and I sincerely hope they will be successful in convincing possible aggressors that war cannot be contemplated.

Even if this is successful, I feel that armies and navies will still not be completely abandoned, for some will be necessary for policing the world to deal with minor situations, but they will be in much smaller measure, and consequently it will be possible to release a considerable number of scientists, technologists and engineers from the study of the means of the destruction of man in order to play their part in the improvement of world conditions. Pride can be taken in the results which have so far sprung from the remarkable co-operation between scientists, industrialists and the Government in war time. It is heartening to find that our Government is continuing to pursue a similar policy in peace time in the field of nuclear physics, the research and development of which are enormously costly and which could not possibly be borne by individual firms.

I must issue a warning that the release of scientists and technologists from activity on war equipment will not be sufficient to solve all the problems to which I have referred in my Address, and The Institution, therefore, must continue to play its part in pressing for still greater facilities, and to continue to advise on the methods of education and training of electro-technical scientists and technologists, in order to overcome the shortage of technical manpower, not only in this country but also in the Commonwealth.

The Institution is to be congratulated on the steps it has taken with industry in producing its film entitled "The Inquiring Mind," with the object of attracting the interest of parents and boys in science and technology, thus stimulating recruitment in these fields. I sincerely hope it will be successful in convincing both parents and boys that the study of science and engineering can be just as powerful in forming well-rounded human minds and will develop at least as good character as the traditional subjects which we know as the Humanities, and are professions of high standing providing stable employment and great scope. It is particularly necessary to attract a great proportion of the young people who still choose to pursue the Arts at various levels, in spite of the need for technologists for the improvement of the world. Another possible source of supply would be by influencing young men who are potentially good, but who take no further steps to carry their education to the graduate standard.

Further interest, I feel, could be stimulated by increasing substantially, with necessary supporting facilities, the numbers of well-qualified teachers in mathematics and science subjects in secondary schools; also by arranging that effective steps are organized continuously and clearly, explaining to boys who have reached the General Certificate of Education level the opportunities of a career in scientific rather than non-scientific subjects, so influencing the transfer of promising boys from one side to the other.

Parents' interest could be stimulated by readjustment upwards of the facilities for awarding scholarships, particularly to help parents who have several children to educate. It is important that technical colleges not affected by the Minister's recent announcement of the upgrading of certain technical colleges to university standard, should not be lowered in standards or facilities. In fact there should be an extension rather than any curtailment. Also, if a student following a part-time course shows great promise during his development, he should be transferred to a sandwich course or full university course.

But I would emphasize that there is great scope for careers masters at schools who have that rare gift of being able to assess a boy's potential development, and they must be given the facility to push forward boys of special ability and to encourage their enthusiasm for their work.

Fig. 3 indicates the seriousness of the position by contrasting the total university population—all subjects—over the years

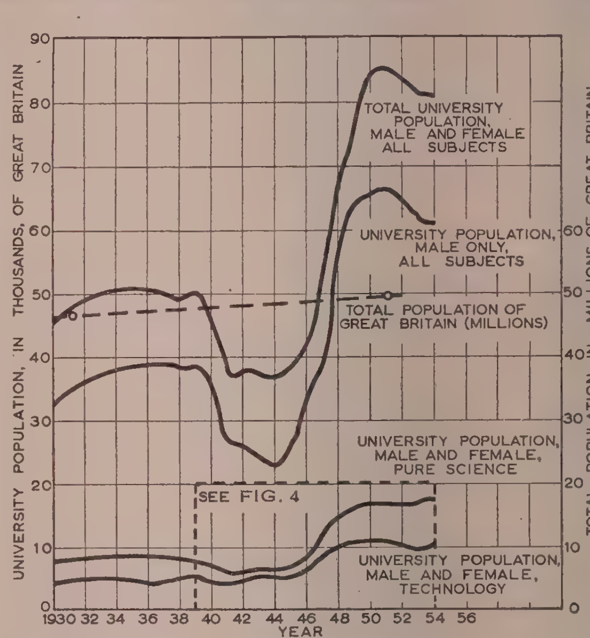


Fig. 3.—Analysis of population of universities in Great Britain.

from 1930 to 1954 with the numbers of students taking full-time courses either in science or in technology (of which the greater part is engineering). Fig. 4 gives, on a larger scale, information of the trend in recent years for pure science and for technology. The curves show the low proportion which science and technology students bear to the total number of university students in Great Britain; also that, even at this low level, technology has not kept pace with pure science, and that in both these cases the post-war increase has levelled off.

As an industrialist who makes a substantial demand on the supply of technical personnel, and who, to help the position generally, has set up very great resources for practical and advanced specialized training of appropriate young men, I have supported wholeheartedly the policy worked out by the Council's expert committees, for I believe they have interpreted correctly the present needs of the whole industry and foreseen properly the future technical demands that will be made on the men concerned some years after their graduation.

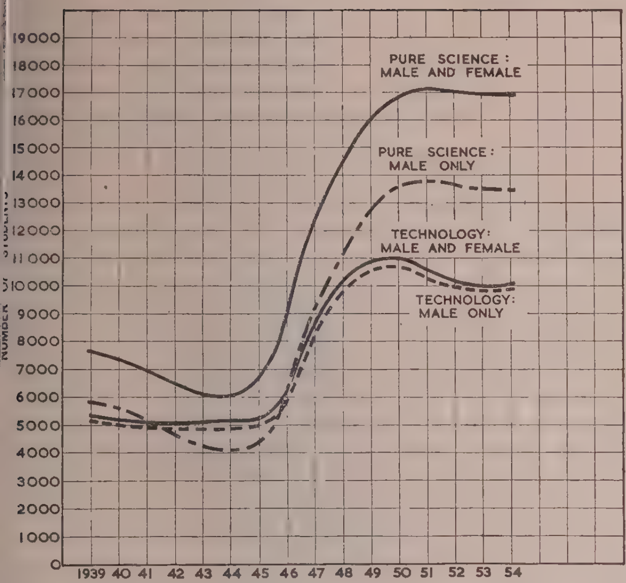


Fig. 4.—Analysis of university population in pure science and technology.

I would say that I support very strongly the warning issued by the Barlow and Percy Committees—that on no account must the demand for quantity be allowed to lower the standards of quality. I would say that remarks have been made to me, both at home and abroad, which implied that some people think that The Institution is raising its educational requirements above what is necessary, implying also that it was limiting recruitment to the profession. I vehemently denied this and pointed out that nothing was further from the truth for, on the contrary, The Institution, while supporting the policy of increased quantity with quality, was pursuing an energetic campaign to attract a larger share of the output of boys from secondary schools to take engineering courses and so increase the input to the technological field and, if of appropriate qualifications, to membership of The Institution.

A very important phase of the supply of technical personnel is the special, great and urgent need for technically qualified men who have the natural gifts for management and highest administration. The ideal men for such appointments undoubtedly are those who are qualified in science and technology and who, in addition, have the inborn qualities of character, powers for leadership and that rarest of all senses—common sense —the combination of which earns the respect and confidence of others.

Such qualities are born in men and are not made by education. The greatest problem is to catch the person as soon as these characteristics appear, and to develop his natural gifts in association with scientific and technical education. I am sorry to say I cannot write a specification for somebody to use for the selection of such men. I can only suggest that the supply could be increased by appointing as school careers masters men who have the gift of spotting them, and who can ensure that they receive appropriate training.

I suggest that the careers masters at schools and colleges with the gift for recognizing those with management qualities can do much to help in this work. The adequate supply of future managers is fundamental, because the best results from design, manufacture and operation can come only with efficient management intimately familiar with the widest aspects of the business concerned.

Great Britain has very few natural resources and without doubt its scientific, technological and skilled people are its greatest economic asset. This being so, they must be employed in the most efficient manner and their use on duplication or further multiplication of effort on the same problems should be avoided, in the interests of the nation and Commonwealth.

Unfortunately, we have not yet reached that degree of rationalization in industry necessary to achieve this important position, which is essential in order to obtain the lowest cost of production, enabling Great Britain to compete successfully in the greatest possible measure in the markets of the world. This it must do to meet the needs of its growing population and of an expansionist world economic policy, so as to achieve for Great Britain the maximum effort in order to pay for the increase in imported materials so necessary if we are to increase our output. Success in these two roles necessitates the importation into Britain of goods and materials which do not exist in this country, and the exportation of further goods to pay for the increase of material so necessary for extending our output to meet increasing world demand.

The deep interest of Members in this matter was apparent to me during my visits as Vice-President to the Centres and Sub-Centres where, in almost every area, questions were raised about the need for more papers on the management and practice of research, manufacture and operation of electrical machinery and on equipment in all fields. I sincerely hope that the Council and the Papers Committee will take note of this, and that Members who can present such papers will respond to this very important request and so add to the continuous improvement in production and operation of electrical equipment.

I feel that, great as the Address of the first President, in 1872, was, at that time there were no fears about the imminent exhaustion of supplies of basic materials. Neither did this question seriously arise as recently as 50 years ago, but to-day the effect of increasing demand and growing population on the world resources of materials has inspired the theme of my Address.

This is not a vague question for the future, but it is with us to-day. If a lowering of the standard of living through the exhaustion of natural resources is to be avoided, immediate action to produce synthetic substitutes of all kinds must be taken now. Electrical engineers will have to play a prominent part in these developments, which, in turn, will bring increased responsibilities to The Institution for ensuring that adequate numbers of qualified electro-technical scientists, technologists and industrial administrators are available for this vital work, which responsibilities will undoubtedly be met.

In conclusion, I would thank you once again for the honour which you have bestowed upon me in electing me as your President, and for your patience with, and attention to, my Address. I would like to pay special tribute to the achievements of the Presidents, Members of the Council, and Secretaries, past and present, who have established The Institution at such a high and valuable level among our national institutions.

Abstract No. 1972 M
Jan. 1956

MEASUREMENT AND CONTROL SECTION: CHAIRMAN'S ADDRESS

By W. BAMFORD, B.Sc., Member.

"MEASUREMENT AND CONTROL"

(ABSTRACT *of Address delivered* 11*th October*, 1955.)

The recent change in name of this Section of The Institution epitomizes the developments which have taken place in the world of instruments during the past 30 years. Measurement and control are the keys to further progress towards that much publicized target, the "automatic factory," and fully automatic production. Enthusiasts have called this a new Industrial Revolution, although there are grave doubts both as to its novelty and its revolutionary effects.

The Measurement and Control Section, however, is inevitably brought face to face with the new problems because of the rise in importance of a new class of measuring apparatus differing in both design and application from the more familiar classes of instruments which were recognized in the original title of the Section. This additional class of measuring equipment is concerned with the application of measurement to control, including the necessary stages of computation.

Popular interest demands that engineers should investigate their responsibilities in this expanding field of automatic production. New words have been coined, and facile forecasts have been made, and it is essential that we should know exactly where we stand as engineers.

A revolution is a great reversal of conditions—a fundamental reconstruction. A revolution always causes upheaval, distress and dissatisfaction. It is, in fact, a most untidy affair, whose participants all suffer discomfort to a greater or lesser degree. This, then, is a markedly strange revolution in which we are taking part, for the evolutionary processes leading to automatic control of production are developing on most encouraging lines. Careful and complete discussions and exchanges of opinion seem to be the antithesis of revolution.

For a moment or two let us look back on the period of our history labelled "The Industrial Revolution." It was preceded by two centuries of development, of scientific discovery and of the spread of knowledge through Western Europe, accompanied by increased demands for manufactured products. The early scientific discoveries led, amongst other things, to the invention of the steam engine—primarily for pumping water out of mines—and then, about the middle of the 18th century, there came an awareness of the new sources of energy which were available, and what were then large sources of controlled mechanical power were put at the disposal of man, who until that time had been dependent on the power of his own hands and feet.

Early in the 18th century there were manufacturing installations but, in general, this was an agricultural civilization, and although there were many employers of labour there were no large single undertakings. The revolutionary changes in social, economic, and industrial conditions which then began were the result of the movement of the power for production from the individual—the craftsman with his hand-operated tools—to the factory, the power-operated mill where the tools were the property of the employer. The resulting mass movement of the rural population to the new urban areas around the factories marked a tremendous change in the lives of these people, and there is no doubt that the change in his living conditions, together with the transfer of his tools from the individual to the authoritatively-controlled group in the factory, had a drastic effect upon character.

The development of automatic control, whether for a single machine, a group of production units, or a complete factory, is a direct evolutionary extension of these historical facts based upon the incentive to research and inquiry which characterized Great Britain in the 18th century. The tendency to concentrate production in larger units is now emphasized—only such large-capacity plants can be adapted to the modern methods of full automatic production. Possibly we shall see a reversal of the moves of population away from the large units as more leisure time is made available; possibly too the individual may begin to use that time in a way which will compensate him for his more mechanical life at the factory. If this surmise is correct, a revolution may be on the way, but we are too near to the scene to give a balanced judgment. We can only hope that such an important social problem will receive due consideration not only from politicians but also from engineers, and that history will provide a kindly title.

The inescapable cost of our astonishing industrial progress, which led to this country's outstanding position in oversea markets is to be seen in the hideous towns of our less attractive industrial areas. Let us hope that other more technical methods of production will reverse this process and correct first errors by providing facilities for good living for all—facilities for working with satisfaction, and relaxing with enjoyment.

It has been said in the United States that, of the total energy expended in industry 100 years ago, human energy provided 22%, machine energy 27% and animal energy 51%. The corresponding figures in 1950 were 2% human energy, 96% machine energy, and 2% animal energy. We must therefore look elsewhere than to a saving in human energy when we discuss automatic controls. It would be far more encouraging to the mass of people if instead of referring to our efforts as "labour saving" we used the term "labour using." Only by using labour fully can we expect the products of that labour to be absorbed.

On the matter of terms and names, I suggest that our age is one of industrial renaissance, and that we should quickly and quietly drop the word "revolution."

When Gutenberg invented printing by movable type in 1442 he accelerated the spread of learning and information by an immense degree. Much scientific research began in that period and the dissemination and interchange of knowledge must have spurred on scientists who would otherwise have been isolated.

To-day, as we have pointed out, the methods of automatic control cannot always effect a saving in manual labour, but they do enable man's control of the machine to be speeded up, i.e. man can transmit his instructions to the machine at a tremendously higher rate—the rate at which the machine can accept them, not the rate at which a man could operate it. The ability to process data, to compute and to control automatically—and also to record such instructions for repetitive production—is a parallel to the period of the renaissance.

Mr. Bamford is with Hilger and Watts, Ltd.

It is interesting also to note that the revival of learning of the 15th and 16th centuries was accompanied by an appreciation of the value of gunpowder in war-like activities, and an interest in exploration by which man found his way round the world, and to places undreamed of. Nowadays we have nuclear weapons and the prospect of space travel.

One of the most repeated—indeed one of the most repeatable—promises of politicians was originally put forward by the Chancellor of the Exchequer some two years ago, namely that the standard of living in this country could be doubled within the next 25 years. What is meant by doubling the standard of living is a debatable question which I do not propose to discuss, but whichever way we look at it, we must either produce more goods, or produce the same amount of goods in less time.

The Chancellor's estimate is one which some economists find it hard to support. The attempt to raise the standard of living by 2% per annum, which is the rate prevailing in the United States, requires a steady rise of 3 to 4% per annum in industrial production and 3% per annum in exports. International trade in manufactured products has, generally speaking, increased at this rate over the past 50 years, apart from the periods of the wars and the slump of the nineteen thirties. Thus, this country would have to maintain its share of world trade (some 20%) in order to achieve its purpose, but unfortunately Great Britain's share has been falling during the past 80 years, and it is now smaller than it was in 1937 despite the major efforts which have been made since 1945.

Supposing we accept the idea outlined by the Chancellor: we immediately find ourselves in another dilemma, because the political policy of full employment and the welfare state ensures that no additional labour can be made available. Mr. Colin Clark, the economist, produced earlier this year a depressing forecast for industrial production, basing his calculations on the size of the labour force. By using the ratio between hours worked and the total national production he arrived at an average increase in productivity per man-hour of only 1½% per annum between 1948 and 1953. As further sources of labour in the country are exhausted the figure of 1½% may prove to be too optimistic for the future, and it seems that we face an imminent and very serious labour shortage.

But this almost completely ignores technical development as a means to increased productivity and it may be that our hope is not so much in new techniques as in the application of the existing techniques in new and hitherto untried situations. This emphasizes the need for the engineer who can apply the techniques rather than the research scientist or inventor to find new ones. Wherever machines can be made to work without attention, or with the minimum of attention, they must be made to do so if we are to follow the plan which has been laid before us.

We cannot afford to neglect the automatic factory or its implications. Our livelihood, to say nothing of our living standard, is absolutely dependent on our maintaining our place as a nation in the industrial and economic struggle. If we, as engineers, refuse to take the path of progress with vigour and enthusiasm, our own living standards, to say nothing of those of the next generation, must deteriorate.

At one time, this country could depend upon the export of raw material—mainly coal—to provide a foundation for our prosperity, even during the time when we undoubtedly led the world in manufactured goods as a result of the Industrial Revolution, which gave us a tremendous margin for a number of years and enabled us to put our standard of living higher than that of any other country in the world.

The picture is now quite different. We are no longer a country which can export raw materials at all. We now import more coal than we export, and we even fire steam locomotives on Polish coal. The industrial progress which was mentioned earlier provided our competitors with means to carry out a number of industrial processes themselves, and with the importance now attaching to economic self-sufficiency amongst the politicians, the barriers against us have been reinforced by protective tariffs and other financial devices. We now find ourselves, not in the van of industrial progress, but somewhat behind the leaders, as we press toward a goal which bears the contemporary label of our improved standard of living.

The economic system is a self-regulating machine, no matter what its political basis may be. But like all self-regulating mechanisms the components and their characteristics must be fully understood and adjusted, otherwise the machine will eventually stop or run away. In either event the effect on the human being is catastrophic, hence the necessity for interference of a transient and spasmodic nature which characterizes our history.

Attempts have been and are being made to study this very complicated mechanism, full of non-linear characteristics of a complex nature—so far without much success. Prof. Tustin* has made notable contributions to this subject, and it is hoped that economists may begin to take advantage of new facilities for computation.

It has been shown how various characteristics of our financial system are related, and that the old swing of boom and depression has been damped down by the application of control forces from time to time. Indeed, the economic life of this country demonstrates the fact that the damping of oscillations increases their frequency while reducing their amplitude. Sufficient is known of the characteristics of our economic system for forecasts to be made, and more than one economist has stated categorically how our present system is likely to behave. Would it not be well worth while to deal with our problems on a more scientific basis of computation? For example, is there a country whose economy is sufficiently simple—I do not suggest the United Kingdom, but probably Australia or New Zealand—for an an analogue computer to be built up? The exploitation of the computer in such a problem would be far less costly in the outcome than the empirical remedies at present employed which all suffer from the same complaints. They are inevitably applied too late, they are based on an error signal which is already incorrect by the time action has taken place, and they take little note of the rate of change of the situation.

Where does the measurement engineer's interest and responsibility lie? There is practically no limit to his field of operations, and even supposing one should limit the discussion to the electrical engineer one finds that sooner or later the majority of measurement and control systems involves electrical engineering problems.

As the use of automatic control increases, so will its opportunities.

In suggesting that scientific investigation into the characteristics of our economic machine would be a valuable enterprise, only one of the non-industrial fields of work was mentioned. There are others with great possibilities of public service.

There is no intrinsic or social value—although there may be virtue—in the making of a measurement; the use to which that measurement is put is the important thing, and I suggest that the best person to apply the measurement or the derived control is an engineer whose training, experience and knowledge give him confidence in his technique, but who has, in addition, absorbed some of the needs of the industrialist—be he chemist, production engineer, or of any other branch. The term "derivative" is often used in a strictly technical sense in discussion between specialists, but do not let us forget that the user of our measurement and

* TUSTIN, A.: "Mechanism of Economic Systems" (Heinemann, 1953).

control techniques is more interested in another derivative—the product of his plant or machines.

Our standard of value must be based on the technical performance of our equipment, and his standard of value on the output of his factory.

One of the American national engineering societies bears a tablet on the wall of its library: "Engineering—the Art of Organizing and Directing Men and of Controlling the Forces and Materials of Nature for the Benefit of the Human Race." The prominence of man in this definition deserves our full acceptance.

The capital cost of providing a job for a man in the United States was stated early this year to be \$13 700. I have been given a corresponding figure for process plant in this country. Modern schemes involve an expenditure of £10 000 per worker, and in some cases several times this amount. In such examples automatic control does not lead to any reduced labour charge, since it is already very small in comparison with the output. It is essential, therefore, if correct and efficient management of capital is to be ensured, that very early in the proceedings initial investigations between research workers, engineers and others should be undertaken, including process details, plant operational characteristics, and the engineering of the plant itself (see Table 1).

The indiscriminate use of the term "automatic control" obscures the real picture to some extent. There are automatically-controlled measuring instruments, machines, processes, and even complete factories. Thus the approach to design of measurement and control equipment must vary in accordance with the scope of the particular project. We can begin to list the functions in which the designer may be involved in normal industrial schemes, and use such a list to stimulate thought on its extension beyond the headings given. Each of these headings implies measurement in one form or another, by dimension or by chemical or physical analysis.

Automatic measurement and control can be applied to the following operations:

Inspection of raw materials and of machined parts.
Machining in all its forms.
Transfer from machine to machine.
Setting by co-ordinates.
Assembly.
Test.
Continuous process of a chemical or physical nature.
Manipulation of dangerous materials, hitherto impossible to handle.

If we assume that the scheme is successful, it follows that the quality of the product is under stricter control than under manual operation, and the limit of quality is dependent upon the integrated errors of the various sections of the process. Being no longer dependent upon human judgment, this improved quality is accompanied by greater consistency as well as the increased production for which we are aiming. We can therefore expect better quality only in so far as we design for it. The plant must be designed for the method, and so must the controls and the product, which leads me to suggest that the primary measuring device should come under close scrutiny in order that it may be of the highest accuracy possible. The electrical instrument manufacturer, for instance, may have to think in different terms from those in B.S. 89 when he proposes to use an electrical movement to control some automatic process.

A particular example will emphasize this. In a distillation column in an oil refinery an improvement in accuracy of control of temperature of 2° C would lead to an increase in productivity of $\frac{1}{2}$% because of the greater accuracy of actual product control. In a modern plant this would represent an increase in revenue of from £100 to £200 per day—roughly £50 000 per annum. Quite enough, one would say, to recompense any investigation into the design of the basic temperature-measuring equipment.

How is this increased accuracy to be achieved? Are instrument manufacturers as satisfied with the mechanical design of their instruments as they are with the purely electrical or electronic sections of their apparatus?

One can also see that the improved quality of the product will have an effect on primary measuring units which will be improved in order to keep in step. New points of design will have to be considered, in particular the speed of response of the measuring device which will have to be related to the equipment under control; components must be of a highly reliable type, and a greater degree of standardization will no doubt be necessary. Integrated plug-in control units of standard dimensions and characteristics are a possibility. The use of new methods of production of control equipment will also have their effect. Above all, co-ordination of design throughout the production unit, be it a machine or a complete plant, is vitally necessary.

If the rate of increase of productivity of both capital and consumer products is even to be maintained, let alone increased, the applications of measurement and control techniques will reach higher levels of complexity, and the need for technical talent will rise in proportion.

Research and development in materials, and the widening limits of their use, must all be introduced beyond the laboratory. Design must be changed and adapted to take advantage of this new knowledge.

Capital expenditure on the new developments is bound to be very high in many cases and we shall no longer be able to depend upon an intuitive, a sentimental or an emotional decision based upon minor experiments. Taking a chance will no longer suffice; there will be far too much at stake. Thus the technical approach assumes a greater importance, particularly in design and product engineering, and in continuous-process schemes. The use of statistical information duly digested by computers must be undertaken.

The value of the technical man is increased, not diminished, by the impact of automatic controls, and doubts have been raised as to the existing ratio between technically trained men and others in industry generally, and in particular plants and factories. It is certainly not high enough for the future.

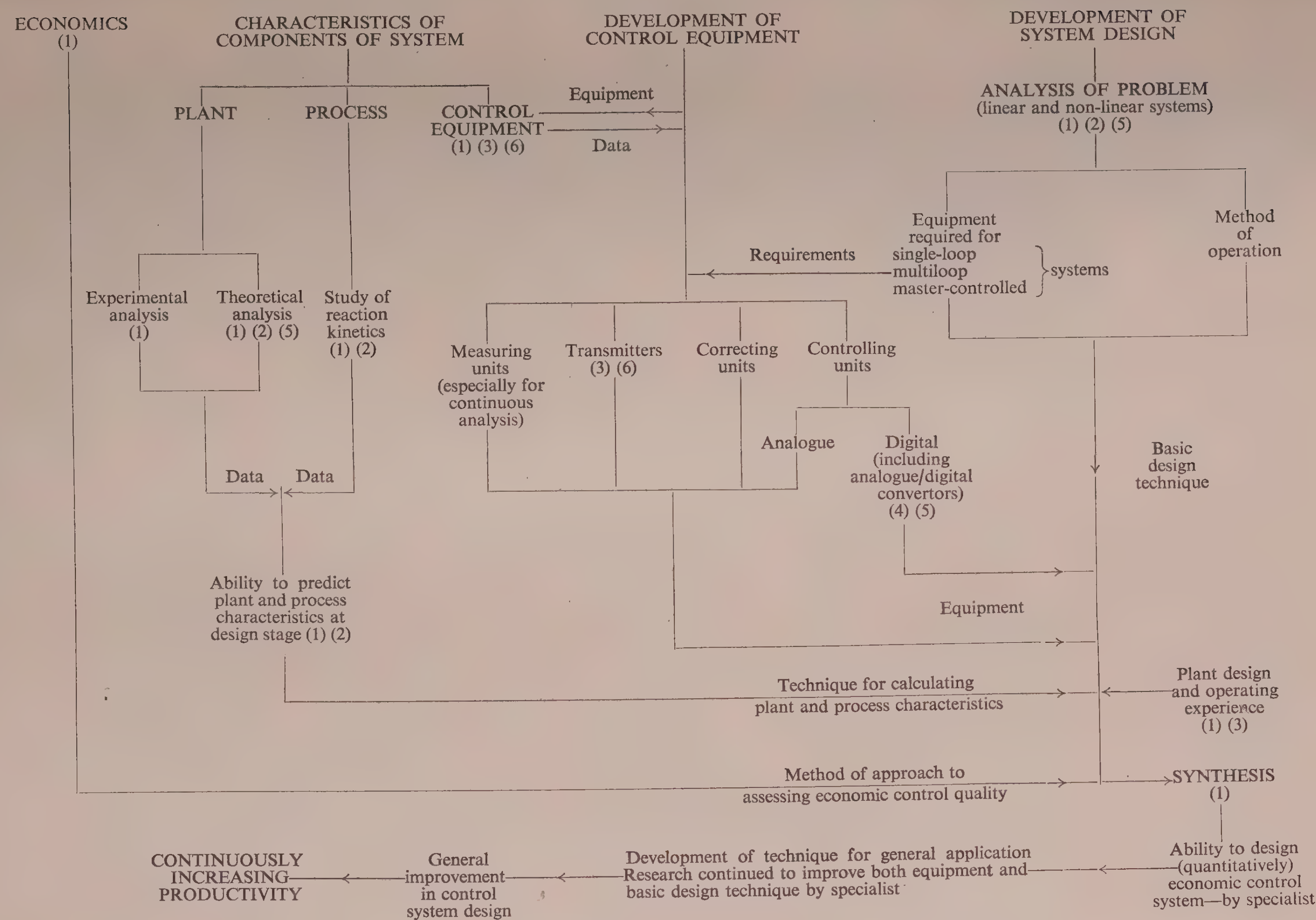
All the evidence at present before us, resulting from conferences, discussions, leading articles and authoritative statements, emphasizes the amount of hard work still necessary before automatic control and automatic production become common. More disturbing, however, is the evidence that trained engineers to design, install and maintain such equipment are likely to be insufficient in numbers for years to come. It is essential that attempts to improve the latter position should be based soundly on fundamental principles, that engineers of the next generation should be trained to think and to assess as a result of true basic understanding of science and mathematics. The pressing need is for engineers concerned with design, production and management, who can make use of the results of research, who can put into useful practice the ideas and principles continually being revealed in papers presented to The Institution and other similar organizations. Invention and discovery must be speedily developed and applied to production. The increasing complexity of many of the concepts demands a trained ability to think and to assess without becoming a specialist in circuit design.

Our approach also to the dissemination of the results of research in the sciences must be adapted to suit the new conditions of speedier change.

The exhibition of instruments of interest to specialists and the reading of specialist papers to similar specialists is not sufficient. Broader dissemination both of technical performance and of

Table 1*

A Proposed Programme for the Development of Automatic Process Control System Design



(1) Users.
(2) University postgraduate research.
(3) Makers.
(4) Makers already in computer field.
(5) N.P.L. Control Division.
(6) B.S.I.R.A.

* This Table is reproduced, by kind permission, from Young, A. J.: "Future Development of Automatic Process Control" (Presidential Address), *Transactions of the Society of Instrument Technology*, 1955, 7, p. 96.

fundamental principles is essential. Specialists must be prepared to interpret their findings to other specialists and to the public.

The rate of progress is so high, and the aggregate of information so great, that the reading of papers and their subsequent publication is likely to cause unacceptable delay; hence the present-day policy of conferences, symposia and similarly organized meetings where integrating papers and others dealing with more isolated matters—all bearing on the same subject—can be introduced and discussed without undue delay. The Institution has, during the past ten years, shown a good example in its organization of such conferences, and next year will occupy a complete week on one of the vital sections of automatic control systems—the digital computer.

Equally, the degree of co-operation between maker and user must be raised. The complete feedback of data on working plant to the maker, together with the derivative of experience by the user—his future proposals—will inevitably improve the sensitivity and control of new design. This is a continuous process with, at best, many inherent time lags, and if it should break down it could bring to a stop any progress.

The presentation, to one or another Institution or Society, of a paper which may be buried among a heap of scarcely digested papers, must not be the end of communication between us. The engineer in industry often has not the time to develop or even to investigate new detailed research or ideas in the theoretical world, but if those ideas are suitably presented, he must and will consider how they may be applied to his own particular problems of manufacture or production.

Above all we must so train our technologists that they do not pass by on the other side when the social consequences of scientific discoveries are considered. Through the engineer our administrators are made aware of the forces which may completely change our way of life and even perhaps our ideals and our hopes. Should we not in this Section of The Institution—which is a nodal point in the modern developments of control and production, and which by definition has the ability to measure and to control not only the product but also increasingly the means of production—be very anxious about such implications of our work?

But, even before we reach that stage, are we quite sure that our theoretical technical conclusions, backed by experiment, are the limit of our responsibility as engineers? Should we not give more time in our deliberations to the actual application of these theories and principles to current problems in order to ensure that, in our own country at any rate, no opportunity is lost to improve our methods in the industrial field. If we as a nation lag behind in this justification of our technical knowledge, our lot is an unattractive one of deterioration and lagging standards of living with unforeseen reactions amongst those who would suffer most in the struggle for survival.

It has been remarked that the dissemination of application engineering information flows more easily in some other countries than in ours, and that material progress is speedier there than it is here. Difficulties of obtaining permission to publish new data based on experience, perhaps due to fear of competition, must be overcome. Can we afford to keep our specialist discoveries, proposals and experiences in isolated compartments for the benefit only of like-minded specialists? There should be a platform where not only the expert may elaborate his theories, but where also the engineer, the man who applies the techniques, may enlarge upon his practical experience for the ultimate improvement of others. It is neither practical nor ethical to put a restraint upon open discussion, and whether we are specialists, scientists, engineers or technicians, means will be found to enable the full reward to the community to be obtained.

Thus, reverting to my earlier remarks, the improvement and speeding up of the transmission of information from man to man will keep pace with the tremendous acceleration of the transmission of information duly digested, from measuring device to the production tool.

The increased confidence which is the reward of the engineer who has successfully evolved or applied a new idea is equalled by the increased confidence which others will have in him when they realize that he is revealing his results for the common benefit. Implicit in this is the idea that the enlargement of our spheres of contact is not a dispersion of our power, but an integration which will distinguish us from the selfish separate communities, which have marked the pages of history with their failures.

In the days of the Industrial Revolution—the mid-eighteenth century—there was no vision of collective human achievement and effort stretching on to the future. The ordinary man and woman lived from day to day—hence the distrust and even hatred of change. Now we have untold opportunities and facilities for comparative education and for constant propaganda, of which every thoughtful man should take advantage both as a learner and as a teacher in so far as his knowledge and experience allow him.

Progress cannot be stayed. New theories, new discoveries and new inventions are only one stage in this process. Another and equally important stage is the application of the technical advance, and that can come about only when, from management down to the tool or process operator, a real appreciation of the objective has been realized. Each man in the chain must obtain sufficient knowledge to satisfy not only his particular technical requirements, but he must also be prepared to consider the economic and social implication of the changes up to the limit of his sphere of activities.

The responsibility of the scientist is particularly great, for he is now providing means of changing the social structure at a greater rate than ever before, and despite the acceptance of these changes as improvements in the standard of life by a large number of people concerned, there is bound to be considerable social adjustment as well as technical improvement.

The flow of technical information continues to increase—the fruit of expenditure on research could hardly be otherwise. It must be presented in a suitable form, above all to the managers, to the technicians, as well as to the engineers, but it must be presented in a form which can produce results. For instance, although it may be very satisfying to present a series of notes or papers on fundamental research on a particular problem, it is far more useful to concentrate the information at some time into a more digested form which can serve for yet another step in progress.

One of our poets, Mr. T. S. Eliot, states this in a much more satisfactory manner than I could hope to achieve, and to quote it will perhaps form a fitting close to my Address:

> Where is the Life we have lost in living?
> Where is the Wisdom we have lost in knowledge?
> Where is the knowledge we have lost in information?

Paper No. 1974 R
Jan. 1956

RADIO AND TELECOMMUNICATION SECTION: CHAIRMAN'S ADDRESS

By H. STANESBY, Member.

"A REVIEW OF LINE AND RADIO-RELAY COMMUNICATION SYSTEMS"

(*Address delivered* 19*th October*, 1955.)

SUMMARY

The evolution of long-distance line-communication systems and the problems associated therewith are outlined. Attention is drawn to the need for certain international standards which facilitate the interconnection of such systems, and the work of the C.C.I.F. in this direction is mentioned. The evolution of radio-relay systems forming parts of internal telecommunication networks is traced, and the salient characteristics of modern systems are summarized. Emphasis is put on the need for international agreement on the preferred characteristics of radio-relay systems to facilitate their interconnection among themselves and with line-communication systems, The work of the C.C.I.R. in this field is mentioned, and the view is expressed that preferred characteristics should be adopted in the near future even though they may need to be reviewed, and perhaps revised and extended, after a number of years have elapsed.

(1) INTRODUCTION

First I wish to say how much I appreciate the honour of becoming Chairman of the Radio and Telecommunication Section, the oldest and the largest specialized Section of our great Institution. I wanted to become a radio engineer many years ago, before the advent of sound broadcasting, and it has been said that the satisfaction of ambitions of one's youth gives rise to the greatest happiness. My own feelings certainly confirm this.

It was my good fortune to be concerned, for some time in my career, with the development of multi-channel-telephone and television-relay systems on coaxial cables, and I found the cross-fertilization of ideas between radio- and line-communication techniques most valuable. So it was with special pleasure that I learned of the Council's decision to extend our Section's activities to line communication. I feel that the decision will be of great benefit to telecommunications as a whole.

For this reason, and because it comes within my own sphere of activities in the Post Office, I shall tell you something of the way in which radio-relay systems have come to be used for television and multi-channel telephony, and of the steps that are being taken to associate them with line systems at home and abroad. But before doing so I shall say something about line systems themselves.

(2) LINE-COMMUNICATION SYSTEMS

In 1871 our Institution came into being as the Society of Telegraph Engineers to cover the first applications of electrical engineering, i.e. the "electric telegraph." In 1876 the invention of the telephone by Bell quickly led to the establishment of telephone systems in the larger towns in various parts of the world. Somewhat later, lines were built to interconnect nearby towns, but as no satisfactory means of amplification was then available they were limited to distances of a few hundred miles. During the First World War, valve amplifiers, i.e. repeaters, were introduced into trunk lines, and extended the distance over which speech could be transmitted so much that in 1915 telephone calls could be made between the east and west coasts of the United States.

In due course, as techniques evolved, it became possible to modulate a carrier with speech, remove the carrier and select one sideband by using a balanced modulator and a filter. This led to the development of carrier systems in which speech sidebands were placed side by side in the frequency spectrum; thus a number of telephone channels could be provided over the same pair of conductors. More recently this technique has come to be known as frequency-division multiplex (f.d.m.) to distinguish it from time-division multiplex (t.d.m.), in which the transmission path between the two terminals of a link is connected in succession to a number of different channels, the whole cycle of operations being repeated again and again at a supersonic rate.

Frequency-division-multiplex telephone systems were applied first to existing overhead lines and underground cables, and made it possible to provide several channels over a pair of wires which formerly handled only one, audio-frequency, channel. However, since the 1930's more ambitious systems have been developed, to provide 12, 24 and up to 60 channels per pair over cables containing 24 or more balanced pairs, and several hundred channels per pair over specially-designed cables containing several coaxial pairs. The provision of large numbers of telephone channels over common paths at frequencies up to 3 000 kc/s or more raised many problems, of which I shall mention only three:

(*a*) How best to assemble and separate the channels at the two ends.
(*b*) How to interconnect similar and dissimilar systems.
(*c*) How to avoid interference between channels traversing the same path.

(2.1) Assembly and Separation of Channels

The superheterodyne principle is used in assembling and separating channels on large-capacity f.d.m. telephone systems; in other words, processes demanding high selectivity are carried out at fixed frequencies, and the translation of the signals to and from the frequencies at which they are transmitted over the line is another operation. (For convenience, the band occupied by an assembly of channels at the terminals of line or radio-relay systems will be called the "baseband.")

In 12-to-60-channel balanced-pair systems the channels are first assembled at 4 kc/s intervals, in groups of twelve channels between 60 and 108 kc/s, using a separate balanced modulator and crystal filter for each channel. As shown in Fig. 1, these groups are then translated, using group modulators and filters, to new positions in the baseband, except for one group which is left in its original position. The process is reversed at the receiving end. Clearly, if each channel had been translated to and from the baseband directly, a greater variety of filters and more stringent filtering requirements would have been involved. A process, similar in principle but carried a stage further, is used for coaxial-cable systems. Referring to Fig. 2, the channels are first assembled in groups of twelve as before, the groups are then

Mr. Stanesby is in the Post Office Engineering Department.

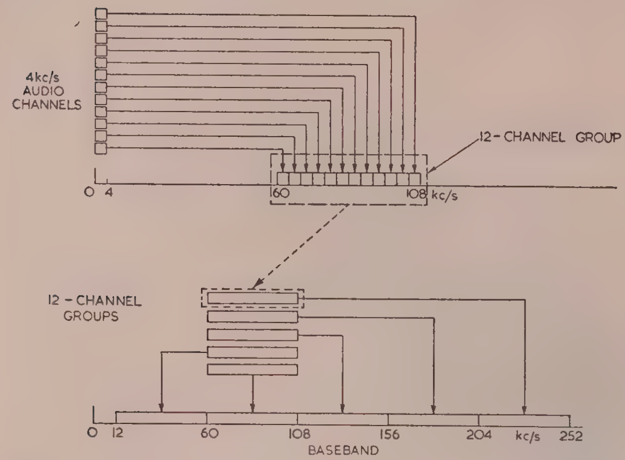


Fig. 1.—Assembly of telephone channels for balanced-pair cable systems.

assembled five at a time to form supergroups which in turn are translated to the baseband.

(2.2) Methods of Interconnection

There is clearly no fundamental problem in interconnecting multi-channel-telephone systems of any type, channel by channel, within these bands. Since the early days of large-capacity multi-channel systems this has been recognized, and in 1938 the C.C.I.F.* first agreed on the channel arrangements for international-junction points in line networks. These arrangements have been adopted widely in national as well as international networks.

(2.3) Intermodulation

When, in a communication system, currents of different frequencies are subject to non-linear distortion, spurious components arise, namely harmonic and intermodulation products, which, if they fall near the original frequencies, can cause interference. Similarly, if a band of frequencies is subject to non-linear distortion, the harmonic and intermodulation products will fall in other bands which may overlap the first and give rise to interference. This is illustrated in Fig. 3, only the second- and third-order products being shown. Higher-order products can arise, of the form $Pf_1 \pm Qf_2 \pm Rf_3 \pm \ldots$, where $P + Q + R$ indicates the order of the product. Interference and noise due to intermodulation constitute some of the major problems in designing f.d.m. systems, and in the line systems already mentioned they are reduced by the following measures:

(*a*) By using highly-linear negative-feedback amplifiers in the main transmission path.

(*b*) By confining frequency-translation processes, as far as possible, to blocks of channels lying within a 2 : 1 frequency ratio, which avoids second-order intermodulation products; and by carrying the translations out at low signal levels.

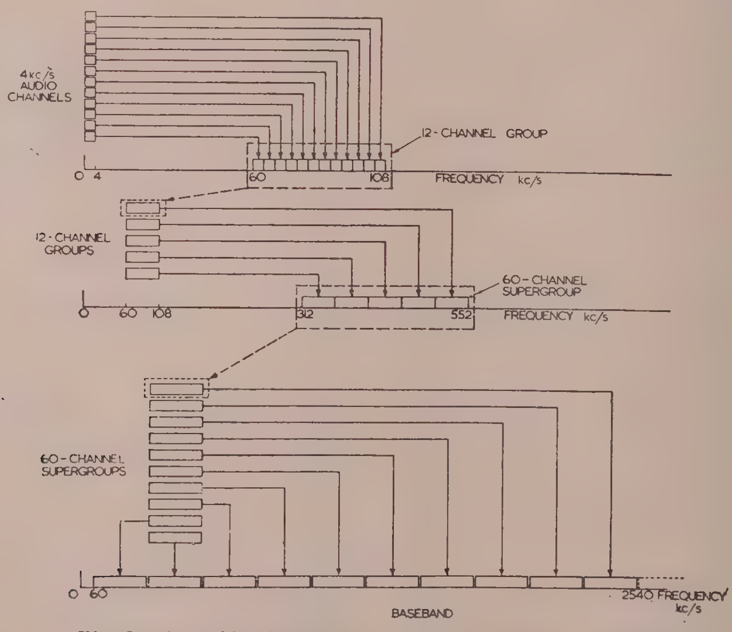


Fig. 2.—Assembly of telephone channels for coaxial-cable systems.

at audio frequencies. But such a method is cumbersome and expensive if many channels are involved, because much equipment is needed for separating the channels and recombining them before and after the interconnection point. In such cases it would be much better to transfer channels in groups of 12 or supergroups of 60 from one system to another, or if appropriate to transfer the channels *en bloc* with no frequency-translation process at all. But to do so there must be agreement on the frequency bands within which the channels shall lie at possible interconnection points, and on the arrangement of channels

These, then, are some of the major considerations affecting the design of f.d.m. line systems—considerations which also have to be borne in mind in designing f.d.m. radio-relay systems.

(2.4) Transmission of Television

When television signals are transmitted over long distances by line, coaxial cable is almost always used. Because coaxial pairs are susceptible to crosstalk and noise below, say, 60 kc/s, the signals have to be raised somewhat in frequency. This is a

* Comité Consultatif International Téléphonique.

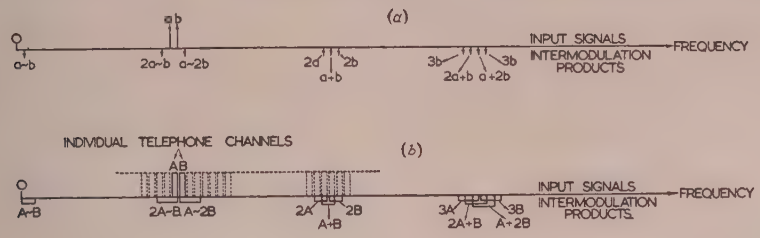


Fig. 3.—Illustration of the way in which intermodulation products can cause interference in multi-channel telephone systems.

(*a*) Second- and third-order intermodulation products for two single-frequency signals.
(*b*) Second- and third-order intermodulation products arising from two channels of a multi-channel system.

complicated process when applied to television, and one which makes the signals more susceptible to non-linear distortion.

(3) TYPES OF RADIO-RELAY SYSTEM

From the beginning, radio has, of course, been used for short- as well as long-distance communication, but not until the 1930's was it used for integral parts of internal telecommunication networks. Initially single-channel telephony links were used, but, as f.d.m. techniques evolved, they were applied to radio as well as lines; and in the United Kingdom many 12- and 24-channel v.h.f.* links are now operating across estuaries and between islands and the mainland.

Contemporary developments may be said to spring from the spectacular advances in u.h.f. and s.f. techniques made from 1940 onwards, notably for radar. Such high frequencies allow broadband systems with highly-directional aerials to be employed, and the signals can be propagated freely over line-of-sight paths with repeater stations at intervals of the order of 30 miles. Two types of system are emerging to-day:

(*a*) Relatively small-capacity systems handling up to 24 channels on a common radio-frequency carrier by t.d.m. methods, and used mainly for private communication systems or in countries where trunk routes are long and lightly loaded. Because they employ time-division multiplex, while line systems use frequency-division multiplex, interconnections between the two are almost always made, channel-by-channel, at audio frequencies, and a number of channels cannot be transferred *en bloc* from one to the other. It is not proposed to consider t.d.m. systems further in this Address.

(*b*) Large-capacity broadband f.d.m. systems in which a radio-frequency carrier is frequency-modulated with a baseband spectrum consisting of a hundred or more telephone channels, or of television. A number of such modulated carriers can be fed through r.f. filters to a common aerial at the sending end, and relayed via repeaters, where they are separated, amplified and recombined, until at the receiving end they are separated, and the baseband signals recovered by demodulation. Because the method of multiplex used is that of multi-channel line systems, these radio-relay systems are potentially capable of being integrated closely with existing networks. They should therefore conform with certain common standards to facilitate their interconnection with line and other radio-relay systems. This subject is of great importance because broadband radio-relay systems are being introduced widely, and is engaging the close attention of the C.C.I.R.†

(4) THE BASIC CHARACTERISTICS OF BROADBAND RADIO-RELAY SYSTEMS

A radio-relay system generally provides one or more broadband channels for each direction of transmission. Where there is need for only one working channel a second may be provided as the most convenient way of guarding against equipment failure; and where two or more working channels are wanted, perhaps for television and f.d.m. telephony, if they have similar characteristics a common standby may suffice.

* V.H.F., u.h.f. and s.f. refer to frequency ranges of 30–300, 300–3 000 and 3 000–30 000 Mc/s respectively.
† Comité Consultatif International des Radiocommunications.

If telephony is transmitted it is desirable to assemble the individual telephone channels in the same way as for coaxial cables, so that radio and cable systems can be readily interconnected. Television, however, is more easily handled if the signals are applied to the radio-relay systems in their original form, whereas for line transmission they must be raised somewhat in frequency. Nevertheless this difference in practice is not unduly troublesome, because, for operational reasons, points of interconnection are often vision-frequency switching points.

(4.1) Choice of Frequency

As their name suggests, broadband radio-relay systems occupy considerable frequency space. At the input and output terminals of such a system, 600 telephone channels might occupy a baseband extending from 0·06 to 2·54 Mc/s, while 405-, 525- and 625-line television signals would extend from zero to approximately 3, 4·25 and 5 Mc/s respectively. Most radio-relay systems employ frequency modulation, and hence, even for low deviation ratios, the modulated carrier might spread over 6 Mc/s or more. If six such broadband channels are provided on the same system, and, as will be apparent later, different frequencies are used for the two directions of transmission, hundreds of megacycles per second of bandwidth will be occupied. Sufficient space can be found only if frequencies of the order of thousands of megacycles per second are used.

Electromagnetic waves in this part of the spectrum are propagated more or less freely over line-of-sight paths, but above, say, 7 000 Mc/s, appreciable absorption is introduced by rain, mist and snow, and at still higher frequencies by the water-vapour and oxygen of the atmosphere. Meteorological anomalies of the type that tend to cause mirages also influence propagation more at the higher frequencies, and small obstructions reflect the waves more effectively and create sharper shadows. It is not sufficient, however, for the line of sight itself to be clear. Any obstruction that might otherwise intercept an appreciable amount of the energy should be sufficiently removed from the line of sight for the direct path to be at least a half-wavelength shorter than the path from one aerial to the other via the obstruction. This condition, known as *first-Fresnel-zone clearance*, should be satisfied, or nearly satisfied, even under the most adverse conditions of atmospheric refraction. There is a further requirement, namely that appreciable reflected components should not be received from outside the first Fresnel zone, since they may impair transmission, particularly on multi-channel telephone systems. In other words, multi-path transmission is undesirable.

At ultra-high and super-frequencies it is possible to obtain very high directivity with aerials of reasonable size, their directivity being expressed as the ratio of the power radiated in, or received from, the desired direction, to that for an isotropic aerial used under the same conditions. The power gain of a

properly designed aerial, whether used for transmission or reception, is directly proportional to its area expressed in terms of wavelength, and hence, for a given absolute area, the gain increases by 6 dB if the frequency is doubled. The relationship between gain and frequency for a 10 ft-diameter paraboloidal-reflector aerial is shown in Fig. 4, with a radiation diagram taken

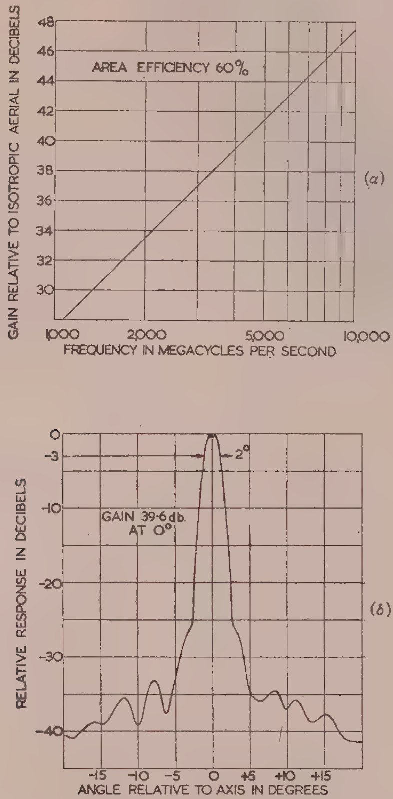


Fig. 4.—Performance of 10 ft-diameter paraboloidal reflector aerial.

(*a*) Variation of gain with frequency.
(*b*) Radiation diagram at 4 000 Mc/s.

at 4 000 Mc/s. High directivity is desirable, not only because it reduces the overall loss between transmitter and receiver, but also because it reduces interference and multi-path propagation.

Neglecting the difficulties of generating and amplifying oscillations at the higher frequencies, which, although at present substantial, are transitory, it is now possible to appreciate the factors affecting the choice of frequencies for radio-relay systems. There must be sufficient frequency space; the higher the frequency used the easier it is to obtain high aerial gains and adequate obstacle clearance, but the more meteorological conditions are likely to affect propagation. Under the Atlantic City Radio Regulations a number of bands may, with certain reservations, be used for radio-relay systems. In the United Kingdom, radio-relay systems have so far been confined to the 1 700–2 300 and the 3 300–4 200 Mc/s bands, but, as these become congested, others, on which work is already being carried out, will no doubt be brought into use.

(4.2) Valves

The generation, and particularly the amplification, of oscillations at frequencies of the order of thousands of megacycles per second presents major problems, and special valves are needed. Grounded-grid triodes can be used, but they must be specially designed with microscopic electrode clearances to reduce electron transit-times, and with the electrodes themselves so shaped that they can form parts of resonant cavities. Moreover, the gain per stage is low, and the alignment of multi-stage amplifiers is difficult, owing to interaction between stages. The alternative is to employ the principle of velocity modulation, in which electron transit-time is an essential and not an undesirable factor. At a point in a uniform electron beam a longitudinally-applied radio-frequency electric field is made to vary the electron velocity with time. Further along the beam these electron-velocity variations build up electron-density variations, i.e. electron bunches, which give up some of their energy to suitably placed electrodes leading to an output circuit. Variations of this theme can be used to realize many different types of valve suitable for generating and amplifying u.h.f. or s.f. oscillations. Two of these, the reflex klystron and the travelling-wave valve, are shown diagrammatically in Fig. 5.

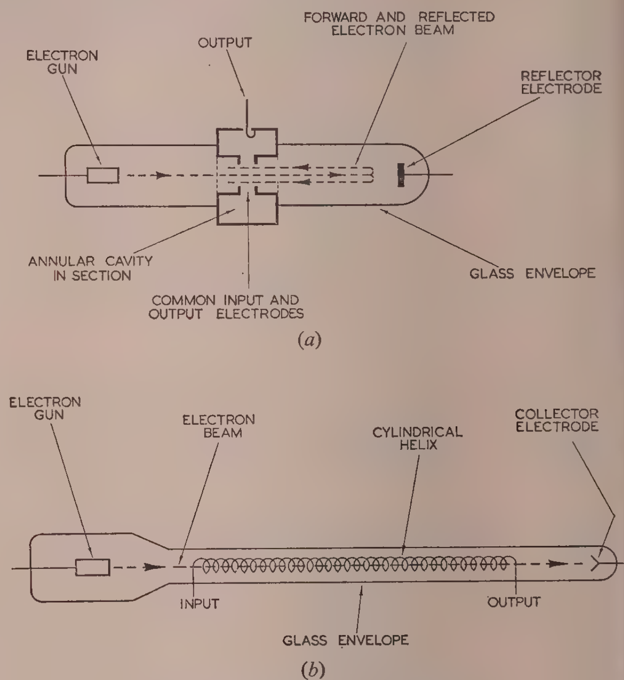


Fig. 5.—Velocity-modulation valves.

(*a*) Reflex klystron.
(*b*) Travelling-wave valve.

In the reflex klystron, widely used as an oscillator, the electron beam is reflected back along itself by a reflector electrode, and the input and output electrodes are common. Velocity variations introduced in the outgoing beam manifest themselves as density variations in the reflected beam, energy from which is used to sustain oscillations in a cavity connected to the electrodes. For amplification, travelling-wave valves are used. They operate by a process of interaction between a wave propagated along a cylindrical helix and an electron beam passing axially through it. In the first part of the helix the beam is velocity modulated, and towards the end it becomes bunched and returns more energy to the helix than was absorbed from it initially; hence there is amplification. As much as 30 dB gain per stage can be obtained over very wide bandwidths, the tuning adjustments are not critical, and output powers of several watts are possible.

(4.3) Distortion and Noise

The main advantages of using frequency modulation instead of amplitude modulation on broadband radio-relay systems are the signal/noise improvement, the fact that amplitude distortion need not be avoided in amplifying the modulated carrier, and the relative ease with which linear modulation characteristics can be obtained. The transmission of television is relatively easy. There is no difficulty in keeping the baseband amplitude/frequency characteristic flat if wide bandwidths are used, and this also ensures that there is little phase distortion. Neither is there difficulty in preserving sufficient linearity; and small departures from perfection only alter the tone gradations of the picture slightly. But for telephony, intermodulation, and hence non-linearity, must be kept at very low levels. Because, with frequency modulation, the baseband signal is transformed into variations of the carrier frequency, the waveform of these variations must be faithfully preserved, otherwise non-linear distortion will result. Frequency is proportional to the rate of change of phase with time, and if a frequency-modulated wave is passed through circuits introducing phase shift which varies with frequency, its instantaneous frequency will be changed slightly. If the phase shift is a non-linear function of frequency it modifies the waveform of the frequency excursions and distorts the baseband signal at the receiving end. Therefore, when multichannel telephony is handled by frequency modulation, the need for linear phase-shift/frequency characteristics is just as pressing as the need for amplitude linearity in line systems, and in both cases departures from linearity cause intermodulation.

In planning radio-relay systems it is important that random noise and, where f.d.m. telephony is involved, intermodulation noise should not exceed tolerable limits. Random noise is reduced relative to the signal by increasing transmitter power, aerial gain and frequency deviation. But intermodulation increases with deviation because a larger frequency excursion is more likely to extend into regions where the r.f. and i.f. phase characteristics are non-linear. Since aerial gain and transmitter power are limited in practice, there is an optimum value for the deviation. This optimum value is clearly dependent on the range of fading to be allowed for and the extent to which multi-path propagation, by introducing phase non-linearity, increases intermodulation. To allow for multi-path propagation is a major difficulty in fixing system parameters.

(5) BROADBAND RADIO-RELAY SYSTEM TECHNIQUES

With this background I will outline some of the techniques that might be used on radio-relay systems providing six broadband channels for television or f.d.m. telephony. Referring to Fig. 6, there are for each direction a transmitting terminal, a number of intermediate repeaters, and a receiving terminal, separate aerials being used for the two directions of transmission. At a transmitting terminal there are six baseband-input circuits leading to six broadband f.m. transmitters, the outputs of which pass through combining filters to a common aerial. At a repeater the r.f. signals corresponding to the various broadband channels are separated, amplified, changed somewhat in frequency, and recombined for transmission over the next section. The frequency is changed to avoid local feedback from the transmitting to the receiving aerial, and to prevent signals from one station being received directly at the next but one station if the sites are more or less in line and propagation conditions are good. At a receiving terminal the r.f. signals are separated for the last time and amplified, and the baseband signals are recovered.

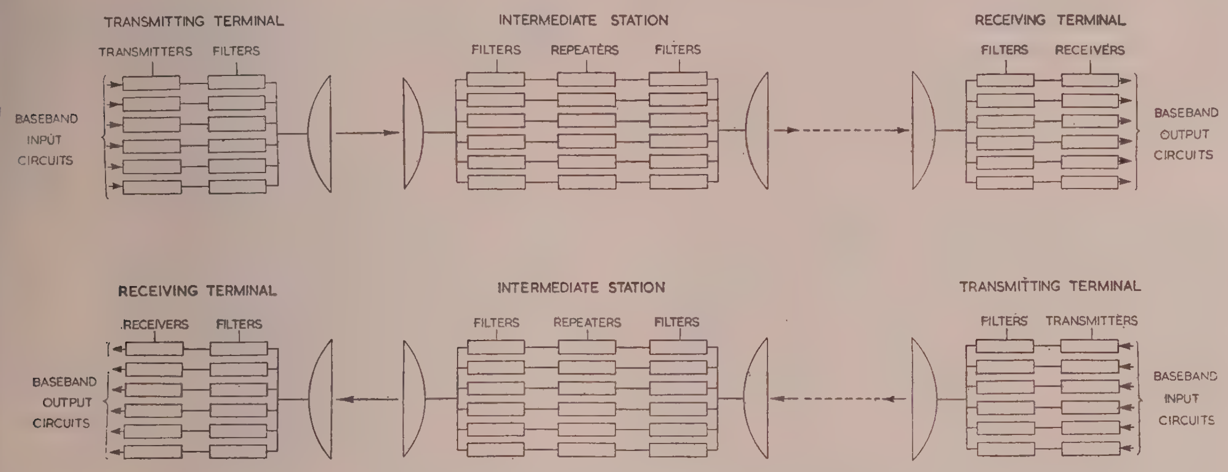


Fig. 6.—Radio-relay system for six broadband channels.

The various processes involved will be considered briefly to illustrate current practice.

(5.1) Generation of Modulated Carrier

The most direct way of frequency-modulating the carrier is to apply the baseband signal to the reflector-electrode of a reflex-klystron oscillator. The baseband power required is negligible, and linear frequency deviations of several megacycles per second are readily obtained. As already mentioned, the deviation used is a compromise between conflicting requirements of low basic noise and low intermodulation. For f.d.m. telephony the deviation for each telephone channel, loaded with standard test tone, is usually between 0·05 and 0·2 Mc/s, the larger value being used on systems carrying many channels. For television the peak-to-peak deviation is usually between 4 and 8 Mc/s.

(5.2) Amplification

Grounded-grid triodes or travelling-wave valves can be used in u.h.f. and s.f. amplifiers, but in the United Kingdom, travelling-wave valves are generally used in large-capacity systems. Hitherto, because suitable low-noise valves could not be obtained, very-low-level u.h.f. and s.f. signals have not been amplified

directly. They have been translated to an intermediate frequency, for which low-noise valves are readily available, and amplified at that frequency. Then, if a repeater station is involved, the signal is raised again in frequency, and amplified to a level of 0·5–20 watts. But if it is desired to recover the baseband signal at a terminal station the signal is limited and demodulated. Fortunately, low-noise travelling-wave valves have recently become available, and, for repeaters, the double-frequency-changing process may soon be unnecessary. Thus repeaters could be made with three or four travelling-wave valves providing all the amplification, the small change in frequency needed before signals were passed to the next repeater section being introduced in one of the travelling-wave-valve stages.

(5.3) Frequency-Changing

The translation of the low-level received signal to an intermediate frequency is carried out in a silicon-crystal frequency-changer. The noise-factor is about 10 dB; in other words, the random-noise output is some 10 dB higher than that due to thermal noise in the circuit connected to the input terminals. For translating an i.f. signal back to ultra-high or super-frequency, however, a germanium-crystal frequency-changer is generally used, because much higher signal levels are involved.

As already foreshadowed, a travelling-wave valve can be used to introduce moderate changes in the frequency of a u.h.f. or s.f. signal. If oscillations of a frequency equal to the desired change are applied to the electron gun of a travelling-wave valve, they vary the beam velocity and phase-modulate any signal amplified by the valve. In this way high-level sidebands can be generated which are spaced on either side of the original signal by the desired frequency change, and one can be selected with a filter. Such a frequency-changer has a gain of 10–20 dB, whereas a crystal frequency-changer has a loss of 10 dB or more.

(5.4) Combination and Separation of Broadband Channels

A typical frequency pattern for a six-broadband radio-relay system is shown in Fig. 7. In any given repeater section six "go"

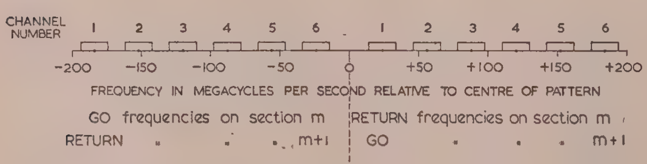


Fig. 7.—Typical frequency pattern for six-broadband radio-relay system.

and six "return" channels are grouped in two adjacent bands, and at a repeater station corresponding "go" and "return" channels are interchanged in frequency. The filters needed for combining and isolating these channels are made up of sections of waveguide, some forming resonant cavities and others forming connecting links and junctions. The filtering is more easily described in terms of reception rather than transmission. From a receiving aerial the combined signals arrive first at a branching filter where most of the signal power of each channel is diverted into a separate branch. In each branch further selectivity is introduced by a waveguide filter consisting of a number of resonant cavities connected in tandem through three-quarter-wavelength sections of waveguide. A four-cavity filter of this type is illustrated in Fig. 8.

A similar arrangement of waveguide filters is used for combining broadband channels, but the direction of transmission is, of course, reversed.

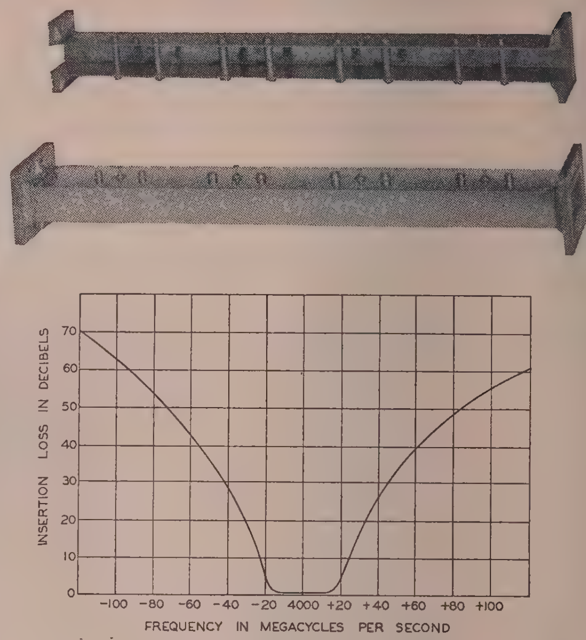


Fig. 8.—Construction and insertion-loss/frequency characteristic of 4 000 Mc/s four-cavity waveguide filter.

(5.5) Aerial Systems and Feeders

Paraboloidal-reflector aerials are widely used in radio-relay systems. In a typical repeater station, these are two 10 ft-diameter aerials facing in each direction, one for transmission and one for reception, and each aerial is connected to the internal equipment by a feeder of rectangular-section waveguide.

All the aerials are similar, but it is convenient to consider their action in terms of transmission. At the focus of the paraboloid the waveguide terminates in a small directional feed which distributes radiation over the surface of the reflector, whence it is reflected in a narrow beam. The feed is designed to irradiate the reflecting surface without allowing appreciable energy to fall outside the periphery, because this would give rise to backward and sideways radiation. In practice such an aerial can be made to have an effective gain equal to that of an ideal uniformly-irradiated paraboloid having only 30–40% less area, combined with sideways and backward radiation which is at least 40 dB below that of the main beam.

(5.6) Interconnection of Radio-Relay Systems

The possibility exists of interconnecting radio-relay systems in a number of different ways.

(*a*) At radio frequencies, extending over one or more broadband channels.

(*b*) At intermediate frequency, one broadband channel at a time.

(*c*) At baseband, one broadband channel at a time.

(*d*) One supergroup at a time.

(*e*) One group at a time.

(*f*) One audio-frequency channel at a time.

Methods (*c*) to (*f*) are, of course, also available for interconnecting radio-relay and line systems. Clearly, if there is agreement on the frequency bands, levels and certain other characteristics at all possible points of interconnection, there is no difficulty in achieving full flexibility in the most economical

ay; but if there is lack of agreement it may be necessary to dopt less economical alternatives.

Radio-frequency interconnection is clearly best when the nds of the radio-relay systems to be connected are miles apart, .g. on opposite sides of the English Channel, and broadband hannels have to be extended from one system to the other. ntermediate-frequency interconnection is best when broadband hannels have to be extended from one radio-relay system to nother sharing the same station; indeed, agreement on the .f. characteristics is a most valuable aid to flexibility. Baseband interconnection must be used in extending broadband hannels from radio-relay to line systems, and from one radio-elay system to another with different i.f. characteristics, but the atter method involves additional noise and distortion owing to lemodulation to, and remodulation from, the baseband, which an be avoided if i.f. interconnection is possible. Interconnections involving one supergroup and one group at a time would be arried out in the bands 312–552 kc/s and 60–108 kc/s respectively, ands which these blocks of channels normally occupy at one tage in the process of assembling or separating f.d.m. telephone hannels. Finally, interconnection at audio frequencies, although he most flexible arrangement, should be used only when operaional conditions demand, because much equipment is needed or separating and recombining the channels.

6) THE WORK OF THE C.C.I.R. ON RADIO-RELAY SYSTEMS

Most long-distance connections involve joining in tandem hannels, or blocks of channels, from a number of different line r radio-relay systems. The need for agreement on certain haracteristics to facilitate such connections has already been nentioned, and this applies especially to international circuits, vhere adjoining systems may be controlled by different adminisrations. Indeed it might well be claimed that agreement at nternational level should precede the establishment of national tandards, otherwise a diversity of the latter might make interational agreement difficult, if not impossible.

The C.C.I.F. is the body that formulates standards for interational line systems, and, as might be expected, its recomnendations affect the planning of national networks as well. t is concerned, not only with the characteristics at interconnection oints, but with overall performance, notably in respect of noise nd transmission quality.

The recent and rapid development of radio-relay systems has nade it necessary for the C.C.I.R. to study their characteristics nd prepare recommendations on their international use. In all natters of common interest, for example in the interconnection f line and radio-relay systems, there is close and fruitful collaoration between the two organizations. Where existing C.C.I.F. ecommendations can be applied without change to radioelay systems the C.C.I.R. is adopting them, and where modifications are needed they are introduced. But many other matters, peculiar to radio-relay systems, are being studied intensively by the C.C.I.R. itself.

The C.C.I.R. has plenary meetings every three years, and the next will be held in Warsaw in 1956, but the urgency of the radio-relay system work is such that an interim meeting of Study Group IX was held in Geneva last autumn. At this meeting it was found possible, albeit tentatively, to adopt preferred values for most of the parameters affecting the interconnection of radio-relay systems, and their connection to line systems. Although these characteristics will be reviewed, and perhaps revised, before they are submitted to the Plenary Assembly next year, some, which bear closely on the subject of this Address, may be of interest:

> A preference was expressed for radio-relay systems carrying 24, 60, 120, 240 and 600 telephone channels per broadband channel.
>
> Agreement was reached on the preferred radio-frequency patterns for an assembly of large-capacity broadband channels, and on certain other characteristics affecting radio-frequency interconnections, e.g. the wave polarization and the frequency deviation.
>
> It was agreed that the intermediate-frequency band should preferably be centred on 70 Mc/s, and that certain values be adopted for the impedances and signal levels at intermediate-frequency interconnection points.
>
> Finally, a preference was expressed for certain arrangements of telephone channels in the baseband, and the impedances and signal levels at baseband-interconnection points were specified. Some of the characteristics adopted had previously been incorporated in C.C.I.F. recommendations for line systems, and the remainder were shortly afterwards used as a basis for C.C.I.F. recommendations —a good example of C.C.I.R.-C.C.I.F. co-operation.

If radio-relay systems are to play a big part in the development of international and national telecommunication networks, manufacturers and administrations alike must be given some idea of the characteristics that are preferred. It could be claimed that the time is not ripe to finalize standards for equipment that is still in the formative stage. This may well be true. But I believe it is already desirable to adopt a limited number of preferred characteristics, with the intention of reviewing, and if necessary revising and extending them, in a few years' time. As telecommunication engineers, we should not allow ourselves to be ruled by standards; rather should we make them our servants.

(7) ACKNOWLEDGMENTS

I have covered a wide field in which many have worked, and it gives me great pleasure to acknowledge the achievements of our colleagues, abroad and at home, in the telecommunication industry and in the Post Office. I should also like to thank them for their unfailing co-operation in our joint labours, and to express my appreciation of Mr. W. J. Bray's help in preparing this Address.

CENTRE AND SUB-CENTRE CHAIRMEN'S ADDRESSES

Abstract No. 1978
Jan. 1956

NORTH-WESTERN CENTRE: CHAIRMAN'S ADDRESS

By G. V. SADLER, Member.

"ETHICAL ASPECTS OF THE ELECTRICAL ENGINEERING PROFESSION"

(ABSTRACT *of Address delivered at* MANCHESTER, *4th October* 1955.)

Although traditionally the Address from the Chair is based on the work on which the Chairman is principally engaged, he is also allowed to say what he likes. I propose, therefore, to exercise this latter prerogative, and depart from tradition.

An address on the specialized nature of my own work would be, I consider, of more interest to an individual Section or Group, whereas, gathered in this room are representatives of all branches of the electrical industry in the north-west. We are all here as members of The Institution whose sole reason for existence is, as its motto aptly puts it, to learn and to teach. By virtue of our membership, we ought also to be better electrical engineers, not necessarily in the technical sense, but as men amongst men, and members of an engineering community.

Membership of The Institution brings its duties and responsibilities, as well as its privileges. Our duties are known well enough. Whether we carry them out or not is a matter for individual conscience.

Our responsibilities are less well understood. We, as engineers whose tasks involve us in a multiplicity of details, administrative and technical, have little time to sit back and consider the impact of technological development on other people, on other less developed nations, and even on the present conflict of ideologies in the world. Every new problem, once it is solved, produces another one, and yet we cannot assume that there must be a technical solution to every problem. The human element must come into it somewhere. The creation of wealth by continuous human endeavour is laudable enough in itself, but performed at the expense of human freedom and liberty, it is a travesty of human nature.

We cheerfully, it seems, sacrifice our liberty in the cause of technological progress. But where is it leading us, and how can we measure it? It is vital for engineers to measure the progress of their work by some known standard and take notice of these measurements, or eventually it will engulf them by creating problems impossible of solution by their own technology.

We can, of course, measure our progress in terms of output of available power per head of world population, or any other material form of measurement you desire, but that is not a measure of the progress of civilization, which, after all, is the only kind of progress which counts. As we all know, the real progress of civilization is not just a matter of mind and body; it is also concerned with things of the spirit as well. When we delve into the history and development of scientific thought, and its handmaid engineering practice, we find that the sciences have been closely related to culture and religion throughout the ages. To enable us to see the picture in this country, as it presents itself to-day, I propose briefly to highlight some of the events which have gone into its making, and in so doing, I must acknowledge having drawn on several sources of information which are listed below. It is sometimes assumed that modern science and engineering commenced during the 17th century, and that a good deal of loose and disconnected thinking previous to that time suddenly became rational and ordered. In point of fact, science has roots and ordered progress going far back into the years we now call B.C.

It is a widespread lack of appreciation of this fact which tends to regard scientific development as some isolated phenomenon, and causes it to appear to be so widely separated from other human achievements. All scientific thought does, of course, spring from early ideas and thoughts about nature and the natural progress of the world in which we live, and the Greeks, with their philosophies, exerted a very strong influence on human thought for centuries.

For centuries, scientific advancement was fostered by the Church, but from the 17th century onwards, science pursued its own course, although man's spiritual progress appeared at one time to be linked to scientific progress, through a common belief in the inseparability of both. In fact, the very course taken by science itself destroyed the age-long conception that science was a religious duty, and a part of the worship of God.

We should, however, try to see the background clearly before studying the present position as it affects the electrical engineer. His is the newest established profession, and it is also the one which enters into normal day-to-day life most actively. There is no implied slight here on the civil, mechanical or chemical engineer. The activeness of electricity is apparent to many a man and woman from the moment the electric alarm clock wakens them from slumber and they switch on the light over the bed, through a day of transport, telephones, radiators, television and the rest, until they are lulled to sleep again by the comfort of an electric blanket.

I spoke earlier of the need for measuring and equating electrical-engineering progress with that of civilization as a whole. What, then, is meant by human progress in the widest sense? We must recognize the human problem in the progress of technology. The western world is basically a Christian community, although I would freely admit that many deny this, or choose to ignore it. But I am speaking to responsible engineers amongst whom I feel sure the percentage of atheism is negligible, and of agnosticism quite small. At the same time, I am fully conscious of the wide variety of our membership, in which race, creed or colour is no bar to being united in a common technical interest which, I submit, must include the progress of humanity. If it does not, the purpose of electrical engineering is lost. Surely, in brief, human progress comprises a gradual refining of human nature over the centuries, a gradual elimination of the bad, and a more positive application of what is good.

So many people to-day regard higher standards of living as a natural right, as though the possession of a plethora of electrical and mechanical aids to comfort and pleasure would automatically engender wisdom, generosity and feelings of goodwill to one's neighbour. I am convinced that we have reached a point in history where those engaged in electrical engineering must take

Mr. Sadler is with the Vaughan Crane Co., Ltd.

ock of what they are doing in relation to the general good of ımanity.

In defining our responsibilities, I do not wish to over-nphasize the religious aspect, although, when we turn to history gain, we find that religious influence has been most marked nongst the people of a nation at times of real moral and iaterial progress. It exerts a stabilizing, as well as an inspiring ıfluence, and yet, to me, it does not seem sound to find that echnology has gone so far, and so fast, in the last 100 years, ıat the gulf between it and religion is growing exceptionally ide, and almost a cult of engineering worship is being estab-shed, many fearing or hoping that the engineer alone will etermine the course of the future world. Engineers may well e asked by laymen, "Are the atoms and electrons you deal in eal things?" We can reply in the words of a modern physicist Yes, they are in terms of patterns of events in the physical vorld."

People have become so accustomed to thinking of existence in erms of material things they can see, like books or chairs or tones, that it is difficult to visualize other forms of existence. 'he age-old question "Does God exist?" is primarily one for the heologians to study and interpret. But when we see existence, ıot only in terms of chairs or stones, and men and women, and emember other forms of existence, such as those atoms and lectrons have, I do not think we can reject the idea of still more orms of existence possessed by spiritual beings. These are lifficult questions, but they are contingent with the progress of nankind, and we must help to find the answers.

What can we do about it now? We can argue that the future velfare of mankind does not depend to any special degree on the norals of scientists and engineers. It rests on the morals of nankind as a whole, and that is a matter for ethics, and not for cientists and engineers. Perhaps, but we have our work to do, nd obviously we cannot suddenly turn into teachers. As an xample of what has been done—perhaps unconsciously—owards furthering the progress of man in recent times, witness the rowth of radio and television, in which electrical engineering has layed such a prominent part. We must not forget that both adio and television have indeed, in this and other countries, een the means of bringing many of the best features of art and earning before people in this generation, whose forbears were lenied such great opportunities; and as such it has promoted rogress.

I think that, as members of The Institution, we should give hought to means of bridging the gulf between technology and nan's humanities, thereby increasing his hopes and reducing his lespair. Individually perhaps, there is not a great deal that an ngineer can do on his own. Those who design and administer an do much, as I know many are doing, towards husbanding the naterial resources of the world, which are being so ruthlessly xpended in many ways for gain or profit. When they begin to un out, mankind as a whole may come to resent the manufacture of products designed only to last a few years; therefore they may e replaced, and thereby maintain employment. Mankind may esent the profligate use of materials in spare parts to bolster mperfect designs. The right use of the world's materials, which re not inexhaustible, is incumbent on us as trustees for future enerations, and obligatory for the future well-being and advance-nent of under-developed countries, in which modern technology s being established for the sake of their raw materials, yet ıntapped.

Disce: Doce, the motto of The Institution, is to learn and to teach. I do feel that by acting together on this basis we can further the overall progress of mankind, and set an example to others. The potential of electrical engineering for world good is so high, if only we can ensure that the good news of its benefits is spread abroad, in order to allay misconceptions in the minds of millions of laymen. Their moral progress can be enormously advanced through the impact of material progress, if they are shown how to bridge the gulf between the two.

It would be more than a tragedy if the opprobrium of "smart Alec" ever came to be applied justly to any engineering community or institution, owing to any failure on their part to recognize their cogent responsibilities to those they lead. Looking ahead, are we wise to assume that The Institution of Electrical Engineers is imperishable, or the Institutions of Civil and Mechanical Engineers? The frontiers of engineering have broadened so immeasurably in the last 50 years, and the divisions have so melted into each other, that we must conceive of an Institution of Engineers presently emerging, combining all that is best in the multiplicity of present Institutions. Of specialized sections, there must be plenty covering the everyday needs and activity which is the special work of each of us. Overriding and co-ordinating all this work would be a council of engineers—we might call them engineer philosophers—at whose meetings would be read no specialist papers, but ones in which breadth of vision and knowledge would integrate the myriad engineering developments into their right context, in relation to the needs of man and his progress in the world.

This, then, is the position we have reached to-day. I spoke earlier of the need for measuring the progress of our work, but progress in engineering does not consist in the mere passing of milestones ever faster and faster. If it did, the rest of the world would eventually be left behind, deprived of the incentive to better living which electrical engineers in particular can provide. Real progress must take the rest of the world along as well, to be leavened and encouraged by the technology which, directed aright, can go so far in helping to solve world problems of body and spirit.

It is my privilege to serve you as Chairman for the next twelve months, and if, together, we can kindle a better understanding of our professional interests amongst those who are unsure of them, and place science and engineering in its proper perspective in human affairs and progress, we shall have achieved much. How it can be done is a matter for thought and discussion, but I am convinced we ought to try. It is one way of fulfilling our obligations, and in so doing will, I think, help to draw into our ranks more of the right type of young men for which there is such vital need in the coming years.

Acknowledgments

The author wishes to acknowledge having drawn historical material from the following publications:

HESSE, MARY B.: "Science and the Human Imagination" (SCM. Press Ltd., 1954).

REES, A. H.: "The Faith in England" (Dacre Press, 1941).

METCALF, B. L.: "Chairman's Address: Utilization Section," *Proceedings I.E.E.*, February, 1954 (**101**, Part II, p. 1).

O.E.E.C.: "Processes and Problems of Industrialization in Under-Developed Countries" (H.M. Stationery Office, London).

Abstract No. 1996
Jan. 1956

EAST-ANGLIAN SUB-CENTRE: CHAIRMAN'S ADDRESS

By E. T. NORRIS, Member.

"MAN AS AN ENGINEER"

(ABSTRACT *of Address delivered at* NORWICH, 12*th October*, 1955.)

It is the privilege of an Institution President or Chairman in his Inaugural Address to choose any subject he pleases and to discourse on it in his own way without control of referees or papers committees, and so at least to his own interest and satisfaction.

My choice this evening is a broad study of human effort in the development and progress of electrical engineering as we know it to-day.

I am doing this on the basis of an engineer talking to engineers—not as an appreciation or build-up of scientific and engineering achievement, but on the theme that we are all human beings with human limitations and weaknesses. Some of us have more ability than others, and a few merit the title of genius, but we are all groping more or less in the dark with far more failures than successes—constantly achieving a little progress and equally constantly slipping back a little, yet nevertheless steadily increasing all the time our understanding and control of the forces of nature.

Human effort is the essence of engineering progress, or indeed of any other progress. Its nature has changed in quality and form in the course of this progress. It is convenient at first to consider these changes chronologically; that is, in their natural sequence, and for this purpose they may be divided into three stages.

(i) *Individual and Isolated.*—In the early days of scientific development—200–300 years ago—scientific studies were carried out mostly by isolated individuals. The subjects studied were sensible; I do not mean intelligent or reasonable, but, more literally, studies of phenomena appealing to the senses. This meant that in general they were either tangible or visible phenomena and so spectacular that they fascinated or excited the curiosity of contemporary society. We must remember that there were no technical journals in general circulation. Information was interchanged by personal visits or correspondence or by conversations frequently at second or third hand.

The approach to these subjects of study was naturally analytical rather than synthetic. It was perhaps too crude and primitive for us to call it a scientific approach, but we may generously refer to these early workers as scientists.

(ii) *General but Unco-ordinated.*—Interest in scientific and engineering matters became more widespread and practical application became possible, initiating the industrial era and with it the electrical engineering industry. Machinery became commonplace and began its prime function of saving and assisting human labour. In consequence, local finance and commercial activity became concerned.

The latter part of this period saw the origin of personal firms such as Brush, Ferranti, Parsons—men who laid the foundations of electrical engineering. We may be proud that many of these names are still perpetuated in our industry.

(iii) *Organized and Directed.*—Commercial and industrial activity grew to such a scale that it became of national concern, both technically and financially, leading in short to the engineering world as we know it to-day. This position is of course well known to all of us.

The growth of engineering beyond even ordinary commercial and industrial interest is now so great that it has evoked national control and direction. These are dictated by finance and even political and certainly national interests.

The various sciences contributing to engineering are now so interrelated and interdependent that there becomes need for directorial control. This trend towards large organizations and amalgamations necessitates, instead of the individual worker of early days, a pyramid of various strata. At the top there must be an administrator or body directing the policy of research and development, below that a stratum of administration and supervision, below that delegation to the actual execution of the work, and so on down to the machines that are actually going to do the work. This organization in direction is of course not only essential but beneficial in that it enables vast resources and capacity to be co-ordinated and utilized efficiently.

The complications and complexity of engineering development have thus greatly increased. It has become necessary to break down any major development scheme into teams of workers—each working under a leader and each team contributing its part to the main development. One could not, for example, conceive of a digital computer as a one-man development.

In the early stages, human effort was devoted to saving physical labour and replacing it by machinery. Now that this has been achieved, at least so far as most heavy manual work is concerned, human effort has turned more to saving mental labour and mental alertness. For example, digital computers save calculating labour, whilst automatic control and protection replaces mental alertness. An engineer in control of a power station or a telephone exchange does not now have to continually and periodically inspect every pump or every bearing lubrication or every contact. He can sit at a desk and have full information on the performance of every process and machine brought to him. For this purpose he does not even have to be awake; alarms will take care of that. Automatic control will ensure that all processes are carried out, and automatic supervisory control will not only intimate when anything is not working properly but to an increasing extent will automatically take steps to correct or replace the faulty component.

Mr. Norris is with Ferranti Ltd.

Abstract No. 1984
Jan. 1956

NORTH SCOTLAND SUB-CENTRE: CHAIRMAN'S ADDRESS

By J. KNOX, M.Sc., Associate Member.

"ELECTRICAL INTERFERENCE"

(ABSTRACT *of Address delivered at* DUNDEE, *12th October*, 1955.)

Owing to the exceptional expansion of electrical engineering, the profession is tending to become the sphere of specialists. The development of equipment and techniques may affect adversely aspects of electrical engineering other than those they were particularly designed to serve. It is inherent in this matter of electrical interference that, in general, those equipments using smaller powers and employing higher frequencies suffer most. I propose dealing only with three major aspects of the problem. My aim is to illustrate how trouble arises, to give some indication of the magnitude of the problems, the methods adopted to limit or overcome it, and the legal aspect where it has been necessary to have recourse to statute to control it.

Inductive Interference between Power and Telephone Lines

Inductive interference arises when e.h.v. power lines are built close to and running parallel to telephone lines for long distances. The precautions to be adopted are determined by the Postmaster General under powers conferred by Electricity Supply Acts. Where protection is necessary it usually takes the form of the power authority arranging to limit the earth-fault current that can arise on its system to a figure determined by the physical conditions existing. The aim is to limit the induced voltage in the telephone lines under power-fault conditions to 430 volts. In special cases a relaxation is sometimes permitted up to 650 volts.

Where limitation of the earth-fault current would cause an unreasonable restriction on the operation of the power line, the Post Office usually undertakes the protection.

This is achieved by:

(*a*) Installation of three-electrode gas-discharge tubes at intervals along the telephone lines. The effectiveness of this method depends on satisfactory earthing at each protection point. Achieving this is often difficult, as the main source of trouble usually occurs in narrow Highland valleys where the earth resistivity is high.

(*b*) Sectionalizing the telephone lines by isolating transformers, which is a simple and effective method, but entails the use of a.c. signalling on the telephone route and the provision of special terminal equipment to achieve this. The equipment designed for use in this country uses trains of 50 c/s impulses for dialling and other signals. So far, it has only been used on various routes in Scotland. In one case, in order to limit the number of telephone lines on which protection was necessary, several new telephone exchanges were opened. This shortened the subscribers' lines to such an extent that inductive trouble was avoided. The protection was then provided on the junctions to these exchanges.

Corrosion of Underground Plant

A high proportion of corrosion arises from electrolytic effects. The offending stray currents may be due to earth faults on d.c. power systems, or may arise where earth returns are used on traction systems, or an earth-return signalling system exists.

In the past, protection of telephone cables has been achieved by insulating gaps or by the use of some covering over the lead sheath.

In recent years consideration has been given to forms of cathodic protection. Two types are in use, one using the primary-cell effects produced between magnesium billets, located at intervals along the cable track, and the lead sheath of the cable. The second method utilizes a mains rectifier feeding a current of anything up to 5 amp through a massive earth. The first method is suitable in areas of low earth resistivity, but the second is also practicable where the earth resistivity is high. Also, with the second method, distances of up to several miles from the massive-earth electrode system can be protected. The flow of current in either system makes the cable sheath slightly negative to earth, thus causing any stray currents to drain away via the magnesium anode or the massive earth, each of which acts as a sacrificial earth and is gradually eaten away. Unrestricted use of cathodic protection by one undertaker could cause serious corrosion in other buried plant in the same area. A committee representing interested parties has been set up to agree on the permissible conditions for the employment of such protection.

Radio and Television Interference

Radio and television interference is the aspect of the subject most in the public eye. Interference complaints received by the Post Office rose from over 82 000 in 1949 to over 141 000 in 1954. As radio and television have become such an integral part of our national life, this growth obviously cannot be allowed to go unchecked. Voluntary suppression was not achieving the necessary results, and in 1949 the Wireless Telegraphy Act was passed to permit regulations to be made to control interference. This has been followed by regulations dealing with the interference from ignition systems on motor cars, from small electric motors and from refrigerators.

An examination of trends of interference complaints received by the Post Office over the past five years reveals disturbing evidence that most sources of interference are causing increasing amounts of trouble, and the increase in the number of complaints is not merely a reflection of the increase in the listening and viewing public. The complaints are increasing at a greater rate than the number of radio and television licences. The regulations so far published only touch the fringe of the problem. A complete solution will take many years, and will require the co-operation not only of electrical engineers but of all users of electrical equipment. The elimination of interference is not solely a manufacturing problem, but is also a maintenance one.

Mr. Knox is with the General Post Office.

DISCUSSION ON
"A SHORT MODERN REVIEW OF FUNDAMENTAL ELECTROMAGNETIC THEORY"*

Mr. C. Hargreaves (*communicated*): The equation for the induced e.m.f. is given by the author [eqn. (22)] as

$$\text{e.m.f.} = -\iint\left[\frac{\partial \boldsymbol{B}}{\partial t} - \text{curl}\,(\boldsymbol{u}x\boldsymbol{B})\right]\boldsymbol{da}$$

This equation is relied on by mathematical physicists and is thought to be quite general in its application and equivalent to the eqn. (14), e.m.f. $= -\dfrac{d}{dt}\displaystyle\iint \boldsymbol{B}\boldsymbol{da}$ or e.m.f. $= -\,d\phi/dt$, from which it is derived as stated by the author.

Now, if these expressions are truly equivalent they ought to give the same result in any particular application, including an expanding rectangular circuit supplied by a constant current.

The magnetic field is therefore set up by the current in the circuit itself and not by an external uniform field, as in Fig. 7 and in textbook treatments, which are mainly concerned with making it appear that "flux cutting" gives the same result as the "change of flux" theory. I agree with the author when he says that examples which would prove that this is not always so are generally avoided.

Therefore some consideration should be given to the expanding rectangle which does not lie in a uniform field. If eqns. (22) and (14) are equivalent they ought to give the same algebraic expressions for the induced e.m.f.

I have considered this problem and have made detailed calculations, the result of which are as follows:

(*a*) From e.m.f. $= -\dfrac{d\phi}{dt}$

we get

$$\text{e.m.f.} = -4iv\left[\log_\varepsilon\frac{s}{g} + \log_\varepsilon\frac{\sqrt{(L^2+g^2)}+L}{\sqrt{(L^2+s^2)}+L} + \frac{L^2+s^2}{L} - 1\right]$$

(*b*) From the equation e.m.f. $= -\,i\dfrac{dL}{dt}$ we obtain the same expression as in (*a*).

where L is the inductance of rectangle, shown in Fig. F.

(*c*) The e.m.f. calculated from the rate of flux-cutting, owing to conductor CD moving to the right, is

$$-\,2iv\left[\log_\varepsilon\frac{s}{g} + \log_\varepsilon\frac{\sqrt{(L^2+g^2)}+L}{\sqrt{(L^2+s^2)}+L} + \frac{\sqrt{(L^2+s^2)}}{L} - 1\right]$$

i.e. exactly one-half of the previous result. The minus sign prefixing this expression is to show that this e.m.f., like the previous one, acts in the opposite direction to the current flow. In these expressions

L and s = Dimensions of the rectangle.
$g = 0{\cdot}778\,8r$, where r is the radius of the conductor.
v = Velocity of conductor CD, cm/sec.
and i = Current, c.g.s. electromagnetic units.

The e.m.f. will be given in the same system of units.

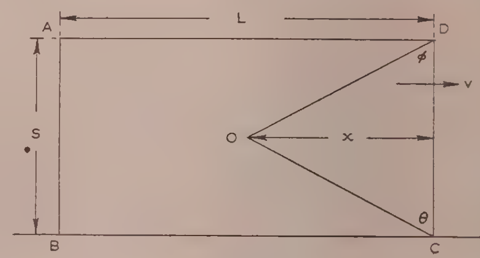


Fig. F

(*d*) We see that if eqn. (22) is to give the same result as eqn. (14), the surface integral of $-\partial\boldsymbol{B}/\partial t$ should give an expression identical to the one given by flux cutting in (*c*). We find, however, that at any point O, inside the rectangle, $\partial\boldsymbol{B}/\partial t$ is negative and given by the equation

$$\frac{\partial \boldsymbol{B}}{\partial t} = -\frac{iv}{x^2}(\cos\phi + \cos\theta)$$

The surface integral of $-\partial\boldsymbol{B}/\partial t$ is given by

$$2iv\left\{\log_\varepsilon\frac{\sqrt{[(L-r)^2+(s-r)^2]}+(L-r)}{\sqrt{[r^2+(s-r)^2]}+r} - \log\frac{\sqrt{[(L-r)^2+r^2]}+(L-r)}{r(1+\sqrt{2})} - \sqrt{\left[1+\left(\frac{s-r}{L-r}\right)_2\right]} + \sqrt{\left[1+\left(\frac{r}{L-r}\right)^2\right]} + \sqrt{\left[1+\left(\frac{s-r}{r}\right)^2\right]} - \sqrt{2}\right\}$$

which is clearly not identical to the expression in (*c*). It should be noted that the surface integral of $-\partial\boldsymbol{B}/\partial t$ gives an e.m.f. which acts in the same direction as the current, thus contradicting Lenz's law.

This is because, inside the rectangle, $\boldsymbol{B}$ is decreasing everywhere as the circuit expands; therefore $\partial\boldsymbol{B}/\partial t$ is negative and $-\partial\boldsymbol{B}/\partial t$ is positive everywhere inside the circuit. Consequently, the surface integral of $-\partial\boldsymbol{B}/\partial t$ is positive; it therefore gives not only the wrong magnitude for the e.m.f. but also the wrong direction.

It appears, therefore, that eqn. (22) is not equivalent to eqn. (14). Furthermore, eqn. (22) cannot be relied upon as a general formula for the calculation of e.m.f.

Mr. P. Hammond (*in reply*): The case of the rectangular circuit discussed by Mr. Hargreaves is full of interest. In this case the magnetic flux arises from the current in the circuit itself, and this circuit is undergoing a change in configuration. It thus becomes essential to include the flux within the material of the conductors in the calculation. It seems to me that Mr. Hargreaves has omitted this flux and this has given him the misleading result that the flux density is decreasing everywhere within the circuit. Actually $\boldsymbol{B}$ is increasing and eqn. (22) gives the correct answer. But it would be fair to admit that eqn. (22) is clumsy in its application to this particular case.

* HAMMOND, P.: Paper No. 1595, December, 1953 (see **101**, Part I, p. 147).

1.314.7 : 537.311.33

Paper No. 1941
Jan. 1956

THE FORTY-SIXTH KELVIN LECTURE

"TRANSISTOR PHYSICS"

By W. SHOCKLEY, B.Sc., Ph.D., Sc.D.

(*Lecture delivered before* THE INSTITUTION, 21*st April*, 1955.)

(1) INTRODUCTION

As an aid to exposition I propose in this lecture to make repeated use of comparisons between the electronics of the conventional vacuum and gaseous variety and the new electronics of transistors and semi-conductors. Conventional electronics is based upon the possibility of producing a satisfactory vacuum. The devices of conventional electronics may be thought of as modifications of the perfect vacuum produced by adding electrons, atoms, ions and various conducting or insulating structures to an otherwise empty space. The ideal or perfect vacuum of conventional electronics corresponds in transistor electronics to a perfect crystal. By a perfect crystal we mean an arrangement of atoms which is geometrically perfect, i.e. the atoms are arranged regularly in rows and planes throughout the specimen so that no atoms are missing from their regular places, no atoms of the wrong kind are present and no atoms are squeezed in places where no atoms belong. Furthermore, the electronic or valence structure of the crystal must also be perfect, so that no electrons are missing and there are no extra electrons present anywhere.

On the basis of the modern quantum mechanical theory of solids it is known that a perfect crystal as defined above must fall into one or the other of two broad classes: metals and insulators. For a metal it is possible for the electrons which bind the metal together to flow through the structure. In this lecture I shall not take time to discuss the theory of ordinary metallic conduction and metals will play no role in the discussion, except where they enter in the role of wires or leads to the devices concerned. The insulating crystals fall into several classes, and those that I shall discuss are characterized by having their atoms held together by covalent or electron-pair bonds. Such crystals will be described briefly below in connection with the diamond structure. Other classes of insulating crystal, such as ionic and molecular crystals, will play no role in the present discussion.

The particles such as electrons and ions of conventional electronics have counterparts in transistor electronics. Chemical impurities, which occur in place of normal atoms in a crystal, play the role of atoms in conventional electronics, and disturbances of the electronic structure of the crystal play the role of the mobile electrons of conventional electronics. The variety of particles is somewhat larger in semi-conductor electronics, however, and in addition to mobile negative charges corresponding to ordinary electrons there are also relatively mobile positive charges. The situation is also somewhat different regarding ions, and in transistor electronics ions of both signs are equally stable. These ions play roles similar to those of ions in gases by neutralizing space-charges and reducing electric fields; but ions in transistor electronics also play the role of metallic structures which are found in vacuum tubes, and a substantial portion of this lecture will be devoted to describing how these ions play such roles and how the desired structures may be achieved in practice.

From the foregoing it is evident that transistor electronics depends upon the presence of various kinds of disorder in otherwise perfect crystals. In the language of solid-state physics, these types of disorder are referred to as *imperfections*. In order to understand the fundamental processes of transistor physics and transistor electronics, it is necessary to know something about the properties of five basic imperfections. In the first part of this lecture I shall introduce these five basic imperfections and describe their most important attributes. I shall next discuss the technology of controlling these imperfections so as to produce semi-conductor structures and devices. In order to illustrate how the imperfections interact with each other in ways which are useful from the point of view of electronics, I shall discuss the properties and nature of p–n junctions. I shall conclude the lecture by saying something about the new physics which has developed about these elementary imperfections; these developments have been made possible by the improvement of materials and techniques associated with the development of semiconductor technology.

There are many materials which may be used in transistor electronics. In this lecture, however, the emphasis will be placed almost exclusively upon germanium and silicon. The most extensive work carried out up to the present time has been applied to germanium, and it appears probable that silicon will be the next material to undergo extensive development. Of these two elements, germanium is easier to control and to purify than silicon, but silicon has some advantages.

The outstanding advantage of silicon over germanium is that transistors or other semi-conductor devices made of silicon will operate satisfactorily at higher temperatures than corresponding germanium devices, and the fundamental reason for this appears to be that the atoms of silicon are bound together more tightly than those of germanium. There are several consequences of this tighter binding: silicon has a higher melting point; it is thus harder to find crucible materials for processing it, and it is consequently harder to purify; on the other hand, the tighter binding of the atoms in silicon causes its electronic structure to be less susceptible to alteration by heat, and thus it may be operated at higher temperatures. The possibility of high-temperature operation for silicon devices is of great potential importance in many areas and in particular in the very demanding conditions met with in military applications involving guided missiles and aircraft.

In closing this Introduction, I should like to point out that a very large number of individuals and organizations have contributed to the development of the devices and knowledge described in the lecture. It would be an undue burden to both reader and author to give references in such detail as to present an accurate picture of the relative importance of the many contributions. For this reason references are restricted chiefly to acknowledgments of sources from which quoted data were

Dr. Shockley is with the Shockley Division of Beckman Instruments, Inc., California, and was formerly with Bell Telephone Laboratories, Inc.

obtained or to originators or inventors of certain especially important subjects.[1–8]

(2) THE FIVE BASIC IMPERFECTIONS

The perfect vacuum of transistor electronics is illustrated in Fig. 1, showing the arrangement of atoms in the diamond structure; this structure is important, since it is the atomic arrangement of both silicon and germanium. These elements, together with carbon (the element of which diamonds themselves are formed) belong to the fourth column of the periodic table. The atoms of these three elements have the common property of tetravalence, i.e. each atom has four relatively loosely bound electrons which can enter into chemical reactions.

The tetravalence of an atom of carbon, silicon or germanium leads it to form chemical bonds with four neighbours. The atom singled out for attention in Fig. 1 is seen to have four symmetrically placed neighbours, and all atoms in the diamond structure are similarly surrounded. In Fig. 1 nearest neighbouring atoms are represented as being connected together with rods. These rods represent symbolically *electron-pair bonds*, or *covalent bonds*. Each bond contains two electrons, one coming from each of the atoms at the ends of the bond.

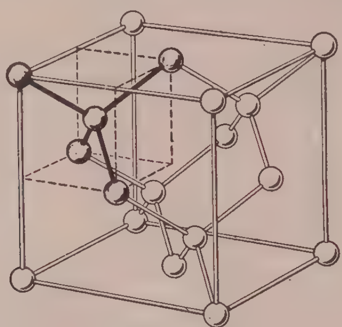


Fig. 1.—The arrangements of atoms in the diamond structure, showing how each atom is surrounded by four neighbours.

A structure like that shown in Fig. 1, in which the atomic arrangement is perfect and all of the electrons are occupied in forming covalent bonds, is an insulator. The electrons in it are inactive and will carry no electric current if an electric field is applied. In order to cause such a crystal to become conducting it is necessary to introduce imperfection into the electronic structure. This can occur in a variety of ways.

Fig. 2 represents the first way that we shall consider of introducing imperfections into the electronic structure of a silicon or a germanium crystal. In this case, a quantum or unit of light energy, known as a *photon*, is represented as entering the crystal. If the photon has sufficient energy, it can break one of the covalent bonds by ejecting an electron and expelling it to some other place in the crystal. In its new surroundings the electron is an *excess electron* over and above those necessary to complete the valence-bond structure, and it thus contributes a negative charge, equal to the electronic charge, to its neighbourhood. Left behind is an incomplete covalent bond. In the language of semi-conductor physics such an incomplete bond is called a *hole*. It is evident that the neighbourhood of a hole contains a net positive charge, since it was neutral before one negative electron was removed from that region.

In subsequent illustrations a minus sign will be used for the excess electron and a positive sign for the hole, in recognition of the charges produced by these imperfections. In general, there is no occasion in transistor electronics to refer to the valence electrons which are in perfect bonds or to the more tightly bound electron in the cores of the atoms concerned. For this reason, no confusion results from referring to an excess electron simply as an "electron." The densities of these imperfections in number per cubic centimetre will be represented by the symbols n_e and n_h (as shown in Table 1) for electrons and holes respectively. The other imperfections are discussed later in this Section.

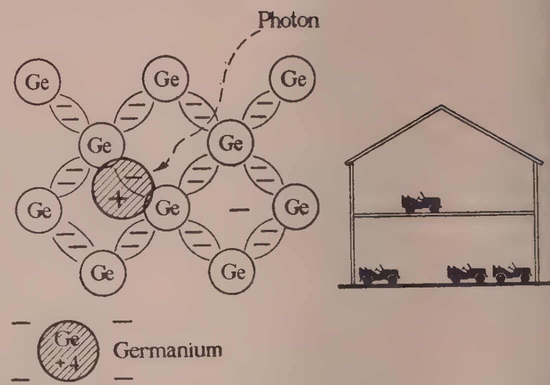


Fig. 2.—The absorption of a photon (a quantum of light energy, represented by the wavy line) may eject an electron from an electron-pair bond, raise it to a higher energy state where it becomes an excess electron: the incomplete bond is called a hole.

Table 1

THE FIVE BASIC IMPERFECTIONS

Name	Symbol	Density
		cm^{-3}
(Excess) electron	−	n_e
Hole	+	n_h
Deathnium	□	—
Donor	⊕	n_d
Acceptor	⊖	n_a

In the physics of semi-conductors some of the most fundamental problems concern the detailed behaviour of holes and electrons. It is not practical in the scope of the lecture to describe fully the theoretical understanding of the behaviour of holes and electrons. I shall, however, return briefly to this topic at the close of the lecture, and indicate the great progress which has been made, both theoretically and experimentally, on this subject.

For the purpose of the lecture, holes and electrons should be thought of as particles and their behaviour regarded as similar to that of electrons in a gas discharge. Thus, the excess electron moves through the crystal with a characteristic Brownian motion, having a certain mean free path and suffering collisions with ions in the crystal. Its motion is also affected by something that is not present in a gas discharge tube, namely thermal vibrations in the crystal. The thermal agitation or heat motion of the atoms of the crystal structure can jostle the electron and thus limit its mean free path—a process which contributes to the electrical resistance in semi-conductors.

The random motion produced by the jostling of thermal vibrations causes electrons to diffuse. If a group of electrons were released in a germanium crystal at room temperature, they would spread throughout the crystal as a result of random motion. If an electric field were applied to the crystal a steady force would be superimposed on the random motion and the electrons would drift through the crystal in the form of a sort of

ectronic wind. This wind corresponds to a flow of electrical arge and so to a current. Thus excess electrons are a means carrying current in a crystal.

The words "diffusion" and "drift" are used to describe the ndom motion due to thermal agitation and the coherent motion ie to the electric field. Under most conditions the motions are lditive, so that the overall motion is made up of the sum of e two effects. Drift motion is characterized quantitatively by e *mobility*. The drift velocity is directly proportional to the ectric field, the proportionality constant being the mobility. or example, electrons in germanium at room temperature in a eld of 1 volt/cm drift with a speed of about 3 900 cm/sec, so at the mobility is 3 900 cm/sec/volt/cm or 3 900 cm^2/volt-sec. he diffusion constant is measured in square centimetres per cond and turns out to be simply related to the mobility. [It is, fact, given by the relationship diffusion constant = mobility $\times$ T (expressed in volts per electron); a discussion of this eoretically interesting topic will be found in many references.] he fact that the drift velocity is directly proportional to the lectric field is the basis of Ohm's law, since it leads to the result at current is directly proportional to voltage. At high electric elds the proportionality no longer holds in germanium and hm's law fails—another interesting topic of transistor physics hich lies beyond the scope of the lecture.

Although it is no harder to describe the behaviour of a hole an that of an electron, it is much harder to visualize the echanisms producing the motion of a hole. When an electron excited from its normal state in the bond structure to a higher nergy state in the crystal, as represented schematically in Fig. 2, here is a hole left behind in the lower energy states. This hole an move through the crystal also. Experience which a motorist as had with corresponding hole motion in dense traffic suggests at, although the hole can move by a sort of replacement rocess, it can move far less freely than can an excess electron. he analogy between a hole and space in traffic proves to be isleading, however, and both theory and experiment indicate at holes in semi-conductors are nearly as mobile as electrons. ike excess electrons, they should be thought of as particles ndergoing random or Brownian motion and having a mean ee path somewhat smaller than that of an electron.

If an otherwise perfect and insulating semi-conductor is xposed to light, as shown in Fig. 2, hole-electron pairs will be enerated and the crystal will become conducting. If an electric eld is applied to it, both holes and electrons flow through it; eing positively charged, the holes move in the direction of the lectric field, and the electrons, being negatively charged, move the opposite direction. If the light is extinguished, so that no dditional hole-electron pairs are formed, the photo-conductivity ecays as the electrons recombine with the holes.

It is evident that the process represented in Fig. 2 will tend to crease the conductivity of a silicon or germanium crystal by creasing the number of current carriers present. Such an crease in conductivity is known as *photo-conductivity*. If the ource of illumination is removed, the photo-conductivity will ecay as the hole-electron pairs produced by the light recombine, e electrons dropping into holes and thus producing normal lectron-pair bonds. Although the conductivity decreases after e light is removed, it does not decrease to zero, because there a *mass-action law* that governs the density of holes and electrons.

The mass-action law applicable to holes and electrons is milar to the chemical law which applies to the dissociation of ater into H^+ ions and OH^- ions; in water the product of the ensity of the hydrogen ions and that of the hydroxyl ions is a onstant at a given temperature. In the semi-conductor the nalogous equation is that the density of the electrons times the ensity of the holes is equal to a constant which depends on temperature. As shown in eqn. (1), the constant which depends upon the temperature can be factorized into two terms

$$n_e n_h = 10^{38} \varepsilon^{-W/kT} \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

The first factor depends so slightly upon temperature that the dependence can be disregarded in the lecture. The second factor is strongly dependent upon temperature and enters, not only into eqn. (1), but into many important phenomena in semi-conductors. [For reasons beyond the scope of the lecture, associated with the Pauli exclusion principle, eqn. (1) must be modified for carrier densities higher than 10^{19} cm^{-3}.]

The factor which is highly sensitive to temperature is equal to ε, the base of the Naperian logarithms, raised to a power equal to the ratio of two energies. The numerator in this ratio is the energy, W, required to produce a hole-electron pair and the denominator, kT, is thermal energy. Thermal energy, as is well known, in the theory of statistical mechanics can be defined by virtue of the law of equipartition of energy. The equipartition law states that any elementary particle, no matter what its mass or charge, partakes on the average of the same amount of random thermal motion at a given temperature, and the amount of this random motion is directly proportional to the temperature, T, measured on the absolute scale. The proportionality constant, k, is known as Boltzmann's constant. The entire factor $\varepsilon^{-W/kT}$ is known as the Boltzmann factor. From eqn. (1) it is seen that at low temperatures the exponent becomes a large negative number, so that the product, $n_e n_h$ becomes a small number, whereas at very high temperatures the exponential term approaches unity.

The Boltzmann factor plays a major role in transistor physics and electronics. It resolves the conflict between the tendency of particles to lose energy and descend to the lowest energy states on the one hand and the tendency of thermal agitation to drive particles to higher energies on the other hand. In eqn. (1) the temperature dependence of the Boltzmann factor is emphasized. As the temperature increases, the random motion succeeds in lifting more electrons from their lowest energy states in perfect electron-pair bonds to higher energy states, where they exist as excess electrons. In other cases in the lecture the dependence of the Boltzmann factor upon energy will be emphasized. The first of these we shall consider shortly, when comparing silicon and germanium. In later parts of the lecture we shall consider how the Boltzmann factor can be used to analyse the effects of electric fields. When an electric field is present, an excess electron or a hole has different energies in different parts of the crystal. Sometimes the electric fields exist in conditions of thermal equilibrium; under these conditions the variations in carrier density from place to place depend upon the changes in Boltzmann factor due to the changes in energy. In other cases, externally applied voltages produce changes in energy and corresponding changes in carrier density; these voltage-induced changes produce currents, and the dependence of the currents upon applied voltages can sometimes be expressed very simply in terms of the Boltzmann factors. These applications will be discussed below in connection with p–n junctions and junction transistors.

In order to be able to convert voltage most easily into changes in Boltzmann factors, I shall use the electron-volt as the unit of energy in eqn. (1). The electron-volt is the energy which one electron would acquire in falling through a potential difference of one volt. In terms of electron volts, kT has at room temperature a value of approximately 25 electron-millivolts, or 1/40 eV. It is sometimes convenient to express energies directly in terms of a voltage V. For this purpose the symbol e equal to the electronic charge is introduced. The Boltzmann factor may then be written as

$$\varepsilon^{eV/kT} \equiv \exp(eV/kT) = \exp[V/(1/40)] = \exp 40V \quad . \quad (2)$$

where the value 1/40 applies to room temperature and is kT/e expressed in volts.

The advantage of silicon over germanium at high temperatures may be conveniently explained with the aid of eqn. (1). The values of W are approximately 0·7 eV for germanium and 1·1 eV for silicon. At room temperature, the resulting values for the product $n_e n_h$ are approximately those shown in eqns. (3) and (4), namely

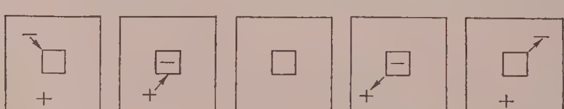
$$n_e n_h = 10^{26} = 10^{38} \times 10^{-12}; \quad W(\text{Ge}) = 0 \cdot 7 \text{ electron-volt} \quad (3)$$

$$n_e n_h = 10^{19} = 10^{38} \times 10^{-19}; \quad W(\text{Si}) = 1 \cdot 1 \text{ electron-volts} \quad (4)$$

In a pure specimen of germanium the only charges present are holes and electrons. The densities of these must be almost exactly equal, for if they differ by a very small percentage, the unbalance of charge will produce enormous electric fields—fields which would promptly produce current flows and neutrality. (Unbalance may exist in very small volumes as discussed below in connection with p–n junctions.) Consequently, taking $n_e = n_h$ we find that both n_e and n_h are about $10^{13}\,\text{cm}^{-3}$ in germanium and about $3 \times 10^9\,\text{cm}^{-3}$ in silicon. Thus, pure silicon will have roughly 3 000 times fewer current carriers than germanium at room temperature, and correspondingly less conductivity. Evidently, for equal carrier densities in pure specimens the absolute temperature of silicon can be higher than that of germanium in the ratio of 1·1 : 0·7. The situation is not so simple as this in transistors, however, but the basic high-temperature superiority of silicon arises from this cause.

The conductivity resulting from the thermal-equilibrium concentration of holes and electrons is known as the *intrinsic conductivity*, since it is an inherent property of the crystal at that temperature and is not produced by such disturbances as light. When the light source which produces photo-conductivity in germanium is eliminated, the increased values of n_e and n_h decay to their thermal-equilibrium values. The law of decay is approximately exponential, so that the density of carriers added by the light decays by a factor of ε in a certain interval of time called the *lifetime*. The lifetime is, in general, not an intrinsic property of the germanium, but depends upon certain additional imperfections, which act as catalysts in the reaction of holes and electrons and may hasten greatly their rate of recombination. For a long time the best-known attribute of this catalyst was its name, *deathnium*. The unknown imperfection was given this name, since it limits the lifetime of the photo-conductivity. It is now known that nickel and copper impurities are very active forms of deathnium and their properties as catalysts for recombination have been extensively studied.

The catalytic action of a nickel centre is represented schematically in Fig. 3. In this diagram the nickel atom or



(a) (b) (c) (d) (e)

Fig. 3.—A deathnium centre in its role of catalysing recombination of a hole and an electron [(a)–(c)] and in generating a hole-electron pair [(c)–(e)].

deathnium centre is represented as a neutral impurity in Fig. 3(*a*). An excess electron wandering through the crystal becomes caught on the nickel atom. The nickel atom now has a negative charge and is in effect a baited trap for catching a hole. If a hole comes near enough to the deathnium centre, the eletron may jump from the centre into the hole, thus producing a normal covalent bond and returning the nickel atom to its neutral condition. It is then ready to catalyse another recombination of a hole-electron pair. By varying the amount of nickel in a germanium crystal one can reduce the lifetime of photo-conductivity by more than a thousandfold. Lifetimes varying from more than 10^{-3} sec to less than 10^{-7} sec can be controlled by the amount of deathnium impurity present. From studies of the statistics of the recombination of holes and electrons, the probability of the centre capturing a hole has been measured and compared with the probability of its capturing the electron. It is found that the ability to capture the hole is about 40 times greater, and this is in agreement with the hypothesis that the centre is either neutral or negatively charged rather than neutral or positively charged. (See Section 8 for some additional details on nickel.)

The deathnium centres will catalyse the approach to equilibrium regardless of the nature of the disturbance. Thus, if by processes that we have not yet discussed the hole and electron densities become depleted, the deathnium centres will produce hole-electron pairs so as to restore the equilibrium values. The process by which a deathnium centre can produce a hole-electron pair is the reverse of that by which it recombines them. This is represented in Figs. 3(*d*) and 3(*e*). In Fig. 3(*d*) the deathnium centre is represented as capturing an electron from an adjacent electron-pair bond, thus producing a hole. The energy required to carry out this process is furnished by heat energy present in the crystal in the form of thermal vibrations. The hole then diffuses away from the centre and subsequently the centre ejects the electron. It is then ready to produce another hole-electron pair or to catalyse recombination.

When the specimen of germanium is in thermal equilibrium, deathnium centres are generating hole-electron pairs at exactly the same rate that they are recombining them. This is an example of the *principle of detailed balancing*. The principle of detailed balancing is a theorem from statistical mechanics which states that under equilibrium conditions every process and its reverse are going on at exactly equal rates.

We must next consider the charge density produced by chemical impurities. As discussed in the Introduction, these play the role of ions in conventional electronics and also the role of structures. The negatively charged deathnium centres discussed in Fig. 3 are examples of such ions, but, in general, deathnium centres are present in such small numbers that their charge density can be neglected. The ions of chief interest in transistor physics are chemical impurities which give rise to *impurity conductivity*, which depends, not on temperature alone, but upon the amount and type of chemical impurity present.

Fig. 4 illustrates the situation produced by the presence of

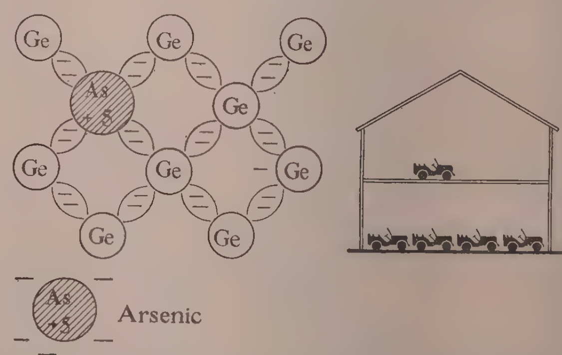


Fig. 4.—An atom from the fifth column of the periodic table can substitute for a normal atom, and lose one of its valence electrons which becomes an excess electron: such an atom is a donor and the resulting crystal is of the n-type.

ipurities from the fifth column of the periodic table. Such ipurities substitute for germanium atoms when a crystal of rmanium is grown from a melt containing them. All elements om the fifth column of the periodic table, such as phosphorus, senic and antimony, act in substantially the same way. Arsenic, r example, has five valence electrons—one too many to fit into e electron-pair bond structure. The extra electron is set free thermal vibrations and wanders about in the crystal. The senic atom left behind thus acquires a charge of $+1$, since s share of the bonding electrons about it amounts to only four ectrons, which is one too few to neutralize the charge of $+5$ n the core of the arsenic atom. An arsenic atom is represented s a plus sign with a circle around it; this distinguishes it from hole and also represents the fact that the arsenic atom is held place by its neighbouring atoms and cannot move through he crystal. An arsenic atom is called a *donor*, since it donates n extra electron to the crystal.

If the density of donors, n_d, is larger than the density of lectrons or holes which would be present in a pure specimen, he conductivity will be due largely to the electrons introduced y the chemical impurity. A specimen whose conductivity arises rom chemical impurities is said to be an *extrinsic* specimen.

The conductivity due to holes in a specimen containing arsenic s even less than it would be in an intrinsic specimen. This is a onsequence of the mass-action law. The extra electrons introluced by donors tend to suppress the hole density. A simple umerical example will serve to illustrate this. Suppose that the ntrinsic specimen contains two electrons and two holes per unit olume, so that the product $n_e \times n_h$ is 4,

.e.
$$n_e n_h = 2 \times 2 = 4 \quad . \quad . \quad . \quad . \quad . \quad (5)$$

f three donors are introduced into the crystal, the condition of lectrical neutrality will be upset. It can be restored, of course, y adding simultaneously three electrons. This, however, would ead to a value of 10 for $n_e n_h$. The correct thermal-equilibrium ondition is found by solving simultaneously for n_e and n_h he mass-action law and the equation for electrical neutrality. The answer so obtained is four electrons and one hole as may e verified as follows: This pair of values satisfies the massction law

$$n_e n_h = 4 \times 1 = 4 \quad . \quad . \quad . \quad . \quad . \quad (6)$$

nd the condition of electrical neutrality

$$n_e - n_h = 4 - 1 = 3 = n_d \quad . \quad . \quad . \quad . \quad (7)$$

ut in words, the last equation states that the charge density of he current carriers (a negative charge density in this case) just alances the *chemical charge density* of the positive donors.

The application of algebra to the mass-action law and to the ondition of electric neutrality shows that there will always be ne particular value each for the hole density and for the electron ensity for any given density of donors. Thus the solution of our electrons and one hole does represent the unique thermalquilibrium condition for the simplified example we have been onsidering.

A crystal of germanium whose conductivity is produced by onors is called *n-type*, because the current carriers in it are redominantly negative excess electrons. Germanium having *-type* conductivity with holes as the predominant current arriers may be made by adding elements from the third column f the periodic table such as boron, aluminium, gallium or ndium. All of these have three valence electrons, which is one o few to complete the covalent bond structure about the npurity atom. The resulting incomplete bond is a hole. hermal agitation will free the hole, which wanders off, thus ntroducing an extra electron to complete the bond structure

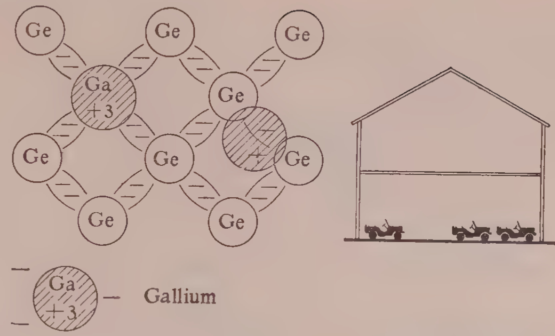


Fig. 5.—An atom from the third column of the periodic table is an acceptor and contributes to *p*-type conductivity.

around the gallium atom, as represented in Fig. 5. The gallium atom thus acquires a negative charge. It is called an *acceptor*, since it accepts an electron from the valence-bond structure. As is represented in Table 1, the acceptor will be shown as a negative sign with a circle around it, again representing the fact that it cannot move, as can a negative electron. The density of acceptors is represented by n_a.

A very important effect occurs when both donors and acceptors are added.[7] One might at first suppose that, if both donors and acceptors are introduced, both electrons and holes would be produced. However, an increase in both n_e and n_h due to chemical impurities would violate the mass-action law. What occurs can be illustrated simply in terms of the example considered above in eqns. (5), (6) and (7). Let us suppose that we have a case in which there are five donors and two acceptors. Under these circumstances the chemical charge density will be the same as for three donors, since the charge on the added acceptors just cancels that on the added donors. Under these conditions, the same values of four for n_e and one for n_h will give rise to a condition of equilibrium in accordance with the mass-action law and will also produce electrical neutrality

$$n_e - n_h = 4 - 1 = 3 = 5 - 2 = n_d - n_a \quad . \quad . \quad (8)$$

Compensation[7] is the name given to the effect of cancelling one chemical charge density by another. It is a very useful effect practically, since it is much easier to add chemical impurities of one kind than to remove those of the other type. We shall discuss in later Sections examples of how this effect may be used in producing semi-conductor devices.

The words *doping*, *majority carrier* and *minority carrier* are frequently convenient aids to describing semi-conductors, and a few examples will show how they are used. A crystal with added arsenic is said to be "arsenic doped" and may be prepared by doping the melt with arsenic. In an *n*-type specimen the electrons are majority carriers and the holes are minority carriers, and the opposite is true in a *p*-type specimen.

In concluding this Section on the five fundamental imperfections I shall discuss one experiment which has now become a classic in the field of transistor physics.[9] This experiment is in essence one which demonstrates the existence and behaviour of holes and permits some of their properties to be measured.[10] The experimental arrangement which was first used for this purpose is illustrated in Fig. 6. The *n*-type crystal of germanium is cut in the form of a rod about 1/32 in in cross-section and approximately 1 in long. A sweeping field is applied from end to end of the rod by a battery, the field acting in such a direction as to draw electrons from right to left down the rod. Any holes present in the rod would drift from left to right.

The contact on the left in the diagram plays a role similar to

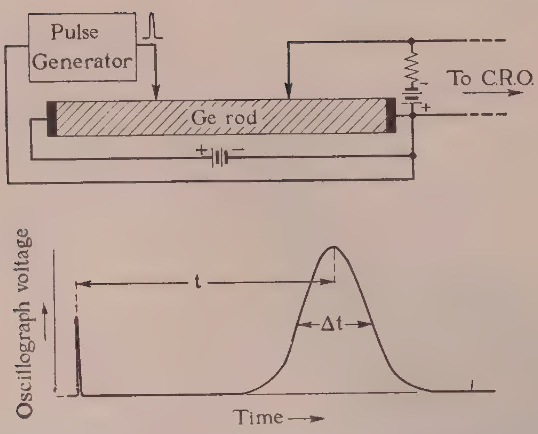


Fig. 6.—The experimental arrangement used for observing the drift of injected holes and measuring their mobility.

an emitter in a point-contact transistor. When the pulse generator attached to it produces a pulse of positive current to the point, holes are injected into the rod. This process can be understood in terms of mechanisms which occur in *p–n* junctions as discussed in the next Section; it is not quite so well understood for a point contact. The holes injected into the rod are then drawn down the rod by the sweeping field. After a time, they arrive in the neighbourhood of the point on the right. This point plays the role of a collector point in a point-contact transistor. A negative bias is applied to it so that it tends to attract holes. When the pulse of holes passes by the collector point, the current drawn by the collector point increases and the voltage across the resistor changes; this change in voltage may be observed with the aid of a cathode-ray oscillograph.

In the experimental arrangement the operation of the pulse generator is accomplished electronically and is synchronized with the functioning of the oscillograph, so that just before the switch is closed the electron beam in the oscillograph starts to move from left to right across the face of the tube. At time t_1 the switch to the emitter point is closed for a brief moment; the time of closing is indicated by a "pick up" signal on the face of the oscillograph. After this nothing happens until time t_2, when some of the holes arrive at the collector point; the concentration of holes builds up for a moment and then decays as the group of holes injected at time t_1 pass the collector point. The arrival pulse at the collector point is not so sharp as the "pick up" pulse, because the holes, which were injected approximately at one point and at the same time, spread out by diffusion, so that by the time the group of holes reaches the collector point it may fill the cross-section of the rod and extend along it several diameters.

It is evident that this experiment permits observation and measurement of both diffusion and drift. It is possible to measure the distance between the points and the electric field between the points; moreover, by calibrating the oscillograph, the time of travel may be measured. Thus the drift velocity may be measured directly, so verifying the fact that the disturbance occurring at the emitter point behaves precisely as would be expected if the emitter point injected small numbers of positive carriers into the rod. For example, if the distance between points is doubled, the time lag between pick-up at t_1 and the arrival of the pulse is also doubled. This result shows that the carriers drift at a constant speed in the rod. But if the sweeping field is doubled, the time lag is cut in half. This fact shows that the speed of the carriers is proportional to the electric field. If the polarity of the sweeping field is reversed, we would expect the injected carriers to be drawn to the left in the filament, so that none arrive at the collector point, and it is found experimentally that this is true.

As was indicated above, the spread of the time of arrival of holes is a measure of the diffusion constant. From studies of the dependence of this spread upon the transit time from emitter to collector, it can be verified that the holes spread out in accordance with the laws expected for diffusion phenomena. The value of the diffusion constant $\mathscr{D}$ can also be measured.

A large number of investigators have performed various experiments of this sort. They have also experimented with the case of electron injection into *p*-type germanium and have dealt with the two corresponding cases for silicon. The best current values[10] of mobility and diffusion constant deduced from these experiments are tabulated in Table 2.

Table 2

MOBILITIES AND DIFFUSION CONSTANTS

	Electrons		Holes	
	μ	$\mathscr{D}$	μ	$\mathscr{D}$
	cm²/volt-sec	cm²/sec	cm²/volt-sec	cm²/sec
Silicon	1 200	30	500	13
Germanium ..	3 900	100	1 900	49

It should be noted from Table 2 that in each case the ratio of diffusion constant to mobility is approximately 1/40, and the dimension of this quantity is in volts; in other words, the ratio $\mathscr{D}/\mu$ is 25 mV. This is simply the quantity kT/e expressed in energy per unit electron charge, as discussed earlier in connection with eqn. (2). The fact that $\mathscr{D}$ and μ are related in this way is expected on the basis of statistical mechanics, and the relationship is commonly known as the *Einstein relationship*. This has recently been investigated in detail for germanium[11] by the means described above. The significance of this value of 25 mV is that an electron moving with random thermal energy will, on the average, be just capable of surmounting a potential hill of 25 mV. In other words, 25 mV is the electric potential corresponding to thermal energy for one electron. Put in another way it can be stated that, if any electron was set in motion with thermal energy in free space against any electric field, the electron would be slowed down by the electric field, and by the time it had moved 25 mV against the field its velocity would be brought to zero and it would start to move in the opposite direction. The fact that a value of 25 mV is obtained shows that the charge of the carriers which are drifting and diffusing in the Haynes experiment is the electronic charge. If it were half or twice this value, for example, the ratio $\mathscr{D}/\mu$ would be 50 or 12·5 mV respectively.

(3) THE *p–n* JUNCTION

Before discussing the means of controlling the chemical structure of semi-conductor materials I shall give a description of a particular *compositional structure*, i.e. a semi-conductor with spatial variations of chemical composition. This structure, the *p–n* junction, is the simplest compositional structure and is the basis of silicon and germanium junction rectifiers and also plays a major role in a variety of transistors.[12,13] In terms of the analogy with the vacuum of conventional electronics, we may imagine that we start with a perfect crystal of germanium. Into this structure we introduce some donors and acceptors and also some deathnium centres; as a result the crystal is divided into two regions, one of which is *p*-type and the other *n*-type.

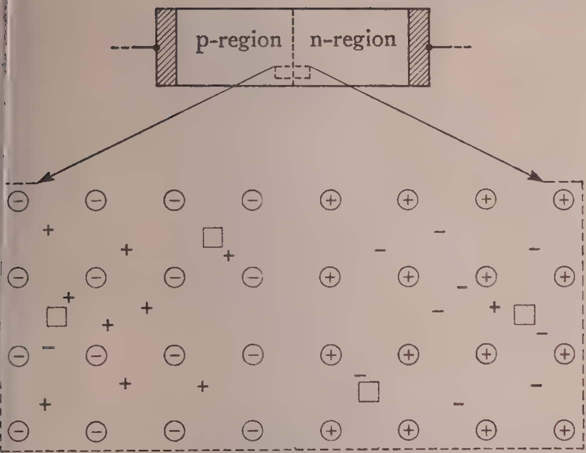


Fig. 7.—A p–n junction and the distribution of imperfections in it.
Symbols are as shown in Table 1.

This situation is represented diagrammatically in Fig. 7; in ı actual case, however, the density of deathnium centres would e much smaller relative to the concentration of donors and cceptors.

If the individual pieces of p-type and n-type germanium ere considered separately, the reasoning presented above would dicate that one is p-type and the other n-type and that each ould be electrically neutral with densities of holes and electrons spectively equal to the densities of acceptors and donors. owever, if a situation of uniform electrical neutrality existed in e specimen shown in Fig. 7, it would not persist. This conusion may be reduced as follows. Since there would be no ectric field present, the holes would spread out by diffusion ıd some of them would enter the n-type region. Similarly, the ectrons would diffuse and some of them would cross the ınction into the p-type region. As a result of diffusion there ould be an excess of electrons in the p-type region and an cess of holes in the n-type region, so that in each region the roduct $n_e n_h$ would be larger than that required by the massction law. This would result in recombination through the eathnium centres and the net result would be that some holes ould be lost from the p-type region and some electrons from the type. This process is illustrated in Fig. 8(b). In other

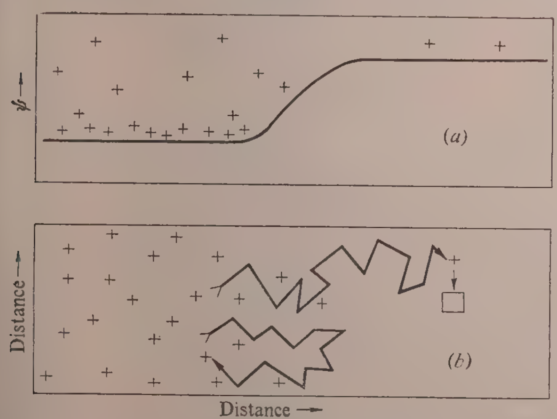


Fig. 8.—Hole current from p-region to n-region in a p–n junction.

(a) Electric potential distribution, with representation of the corresponding Boltzmann distribution of holes.
(b) Typical diffusion paths for holes.

words, there would be a net transfer of negative charge from right to left across the junction. This transfer of charge would cause the p-type region to acquire a negative charge and the n-type region a positive charge; these unbalanced charges constitute a dipole layer which sets up an electric field, which is evidently of such a polarity as to hinder the diffusion of carriers across the junction. When thermal equilibrium is established an electric field will exist at the p–n junction of just sufficient magnitude as to prevent any net flow of current across the junction. As a result an electric field will be present, so that the electric potential will be as shown in Fig. 8(a).

We shall next estimate the increase in electric potential across the p–n junction. Under the condition of thermal equilibrium there will be some holes in the n-type region; since n_e is not infinite, n_h must be greater than zero in order to satisfy the mass-action law. However, this density of holes will be very small if the electron density is large compared with the intrinsic density corresponding to the temperature. Thus the hole density in the n-type region will be less by a large factor than it is in the p-type region. This decrease in density under equilibrium conditions can arise only from a corresponding decrease in the Boltzmann factor. In order to produce this decrease the energy must increase and, consequently, a hole must have a higher potential energy in the n-type region than in p-type region. This increase in potential energy arises from the increase in the potential discussed above. The magnitude of this increase may be calculated as follows: For example, let us suppose that the hole density in the p-type region is 100 times larger than that in an intrinsic specimen and that the same is true for the electron density in the n-type region. Then in the n-type region the hole density will be 100 times smaller than in an intrinsic specimen and consequently 10 000 times smaller than in the p-type region. The Boltzmann factor which will give rise to this difference in hole density may be obtained by solving eqn. (9).

$$10\,000 = \exp\,[V/(1/40)] \quad . \quad . \quad . \quad . \quad (9)$$

The resulting value is 0·23 volt.

For corresponding densities of majority carriers in a silicon p–n junction the height of the potential hill is greater by about 0·4 volt. The reason for this is that, as shown in eqns. (3) and (4), the minority hole density in the n-type region is smaller by a factor of 10^7, because the energy required to produce a hole-electron pair is greater by 0·4 eV in silicon. So that the Boltzmann factor can maintain thermal equilibrium for the hole density across the p–n junction, the hill must accordingly be 0·4 volt higher.

Let us next examine the behaviour of holes and electrons under conditions of thermal equilibrium.

Actually, some holes will diffuse up the potential hill and into the n-type region. These holes combine with electrons, producing a current from left to right in Fig. 8. There is, however, an equal current of holes, which are generated at the deathnium centres, diffuse back to the p–n junction and slide down the potential-energy hill. Fig. 8 represents this situation in a schematic way. The principle of detailed balancing, discussed earlier, states that under conditions of thermal equilibrium there is exact balance between the current climbing the hill and recombining in the n-type region and the generated current originating at the deathnium centres which flows back and down the hill.

The application of voltage to the terminals of the device shown in Fig. 7 destroys the exact balance of the two currents just discussed. In considering the application of voltage we shall neglect any voltage drops at the contacts between the metal electrodes shown in Fig. 7 and the semi-conductors. The effect of the application of voltages upon the currents is represented in Fig. 9. Fig. 9(a)

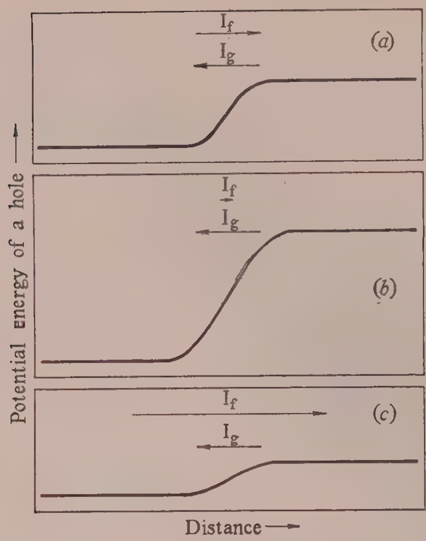


Fig. 9.—Dependence of recombination (forward) and generation hole currents upon applied voltage in a p–n junction.

(a) No applied voltage.
(b) Reverse.
(c) Forward.

shows the thermal-equilibrium condition. The two currents previously discussed are represented by I_f and I_g, these currents standing, respectively, for the current of holes entering the n-region and recombining and the current generated in the n-region and diffusing to the barrier.* For the condition of thermal equilibrium these two currents are equal and opposite. In Fig. 9(b) the situation for a large "reverse" bias is shown. For reverse bias, negative voltage is applied to the p-region and positive voltage to the n-region, so that the electric potential difference between the two regions is increased. If the electric potential is sufficiently high, corresponding to the situation shown in Fig. 9(b), practically no holes can climb the potential hill and I_f drops nearly to zero. This situation is represented by showing I_f as a vector of negligible length, whereas I_g has practically the same value as it has for thermal equilibrium. This independence of current upon bias is referred to as *saturation.*

When forward bias is applied the situation shown in Fig. 9(c) occurs and I_f increases. This increase can be expressed in terms of the change in the Boltzmann factor. In fact, if an applied voltage V decreases the height of the hill by an amount eV compared with its thermal-equilibrium height, the current climbing the hill is increased by $\varepsilon^{eV/kT}$. Since for the thermal-equilibrium case $I_f = I_g$, the value of I_f with applied voltage is

$$I_f = I_g \varepsilon^{eV/kT} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

This gives rise to a total current of holes from p-region to n-region, given by the difference

$$I_f - I_g = I_g(\varepsilon^{eV/kT} - 1) \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

This current is zero when V is zero, increases exponentially to large values for positive V, and decreases to a negative saturation value of I_g when V is negative and substantially larger than kT/e.

Similar reasoning can be applied to the electron current flowing across the junction. The applied potential which lowers the potential barrier for holes evidently lowers it also for electrons; consequently, large electron currents flow under the same voltage conditions that produce large hole currents. In both cases these large currents correspond to flows of minority carriers into each region. In both cases the current is in the sense of transferring positive charge from the p-region to the n-region. In one case this is carried in the form of positive imperfections—holes—moving from p to n, and in the other case it is due to negative imperfections—electrons—moving in the opposite direction. For reverse biases the potential rise is larger and the holes tend to be retained in the p-region and the electrons in the n-region. A saturation current due to generation in both regions flows. If the total saturation current is I_s, the total current for any applied voltage V is given by

$$I = (\varepsilon^{eV/kT} - 1)I_s \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

Evidently I_s is the sum of the two generation currents. This equation is found to be well satisfied for p–n junctions in germanium, and a comparison of the rectification curve as measured with the theoretical formula is given in Fig. 10. It should be

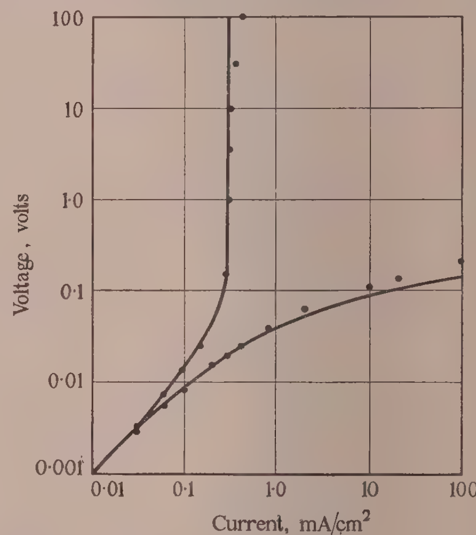


Fig. 10.—Rectification characteristic of a p–n junction.

noted that the separation between the forward and reverse branches of the curves corresponds to a factor of ε when the voltage is $kT/e = 25\,\text{mV}$. This is exactly the factor predicted by eqn. (12). This agreement between theory and experiment is evidence that the imperfections which carry the current in a p–n junction have the charge of the electron. If they had twice this charge, a value of $12{\cdot}5\,\text{mV}$ should be obtained; if they had half the charge, the value should be 50 mV.

For large forward biases the potential barrier between the n-type and the p-type is nearly eliminated. Under these conditions large densities of minority carriers flow across the barrier, and the density of the majority carriers may be substantially disturbed. Although these large signal conditions are of general interest, we shall not consider them further but shall instead proceed with a discussion of the means of producing p–n junctions and other compositional structures.

(4) THE CONTROL OF PURITY AND PRODUCTION OF COMPOSITIONAL STRUCTURES

In conventional electronics the production of the structure is the first step, and this is followed by the production of a vacuum

* The subscripts f and g may be thought of as "forward" and "generation"—a mixed choice which avoids the subscript r, which might be either "reverse" or "recombination." In this notation "forward" is equivalent to "recombination" and "reverse" to "generation."

›ut it. In transistor electronics the process is, in general, ersed and the starting point is the production of pure and fect single crystals. Accordingly, in this Section, I shall cuss first what in transistor electronics is the analogue of the cuum pump. This analogue produces a specimen of germanium which the impurity density actually corresponds to a fair cuum in ordinary electronics: by this means it is possible to oduce germanium in which the density of detectable impurity oms is smaller than the density of germanium atoms by a tor of 10^{10}. The resulting density of about 10^{12} impurity oms per cubic centimetre corresponds to the density of mole- les in a pressure of about 10^{-4} mm Hg.

The most successful semi-conductor vacuum pump is the chnique known as "zone refining."[14–16] The principle of the ocess is as follows: the starting material is "chemically pure" rmanium, which from the electrical point of view is relatively avily doped *n*- or *p*-type germanium with a density of perhaps ^{16} impurity atoms per cubic centimetre. A rod of this ger- anium is put in a crucible and passed through a heating coil hich melts a short length of it. As the germanium is moved, e molten zone passes along the rod. Fortunately for the urpose of this method, with one exception the important donors d acceptors prove to be much more soluble in molten ger- anium than in solid germanium. Thus the impurity atoms tend remain in the molten zone and the material which resolidifies much more pure than the melt from which it solidifies. Thus e molten zone tends to sweep the impurities out of the ger- anium and the purity of the germanium is increased by a factor f 2 or 10 or even more each time a molten zone passes through it. onsequently, if a series of molten zones pass through the ermanium the impurity density decreases in geometric ratio. s was stated above, germanium which contains less than one onor in 10^{10} germanium atoms has been produced by this means.

In order to approach ideal electrical properties it is necessary, ot only to suppress the density of impurity atoms to very low vels, but also to suppress the density of atomic imperfections ue to disorder among the germanium atoms themselves. One f the most common ways in which disorder can arise is for the ermanium ingot to be polycrystalline. A polycrystalline ingot one in which the rows of atoms continue coherently only over ertain limited regions of the body, and in different regions are riented in somewhat different directions. Where these crystal rains or regions of different orientation come together, grain oundaries are formed, and in these boundaries it is impossible or geometrical reasons for each atom to be correctly surrounded y its neighbours. The disruption in the covalent bond structure roduces electrical disturbances which act both as chemical npurities and as recombination centres.

Proper control of the zone-refining process permits the rowing of single crystals at the same time. However, it is requently more convenient to purify the crystal by zone refining nd subsequently grow the refined material into satisfactory ngle crystals in another apparatus.[17] Fig. 11 represents a raphite crucible which may be heated by an induction coil in hich the heating current is controlled by a thermocouple which akes contact with the molten germanium. Crystallization is nitiated by lowering into the molten germanium a small single rystal which has been previously obtained by cutting a specimen rom a polycrystalline mass or from a portion of a previously rown single crystal. In order to grow a single crystal from this eed, the seed crystal is dipped into the melt and the heat supply radually lowered, so that a portion of the melt solidifies on the omewhat cooler seed. In this solidification process germanium toms which attach themselves to the solid germanium fit into e pattern of atomic arrangement in the seed crystal and thus dd plane after plane of atoms all coherently aligned with the

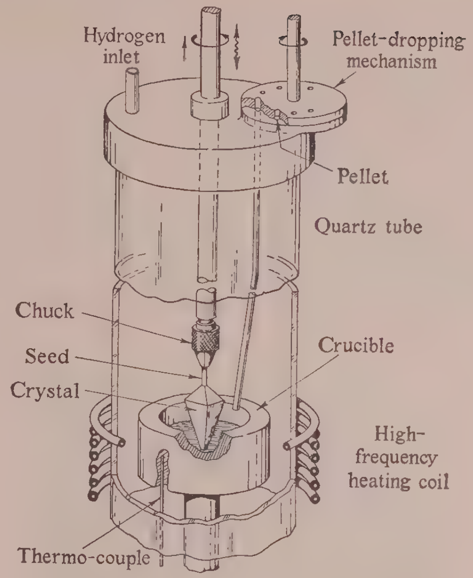


Fig. 11.—Diagrammatic representation of crystal-pulling apparatus.

arrangement in the seed crystal. By careful control of the heat supply and rate of withdrawal of the seed crystal, it is possible to solidify the entire melt into one crystal in which the orientation is determined by the original seed. This process is aided by causing the seed crystal to rotate as well as to move upwards in the solidification process. The rotation brings each part of the growing crystal through the same variations in temperature that may exist in case one side of the melt is heated slightly more than another, and thus causes the crystal to grow with a more symmetrical form.

It is possible to grow a single crystal one inch in diameter and several inches long in 20 or 30 min in such an apparatus, and such a crystal is shown in Fig. 12, the seed crystal having been

Fig. 12.—A germanium crystal grown in the crystal-pulling apparatus.

cut away. It should be noted that the crystal is not perfectly circular in cross-section but shows a tendency to have a square cross-section. This is a consequence of the fact that the diamond structure in which the germanium crystallizes is a cubic structure and the atoms lie on planes parallel to the faces of a cube.

This exhibits itself in the tendency of the crystal in solidifying from the melt to form such faces, so that the crystal assumes a cross-section between a cylinder and a square.

The growing of single crystals of germanium has contributed, not only to the technology of producing devices in transistor electronics, but also to our scientific knowledge of imperfections in crystals. It has been possible to carry out some studies of grain boundaries in germanium which would not have been possible in materials in which the perfection was less thoroughly controlled. As one of these examples we shall consider a so-called small-angle grain boundary which results when an ingot of germanium is composed of two regions in which the planes of atoms are tipped in respect to each other by a very small angle. Such a situation is represented in Fig. 13, the relative

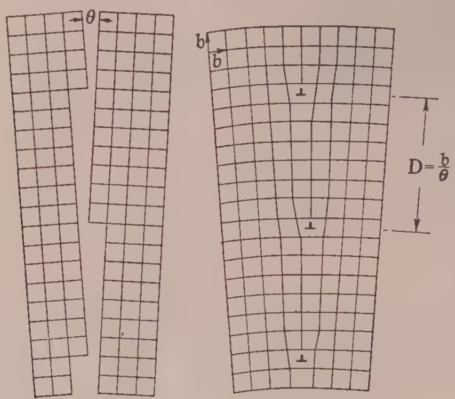


Fig. 13.—Two slightly misaligned crystal grains when contiguous in a polycrystalline mass are joined by a small-angle grain boundary composed of a row of dislocations.

(*a*) Surfaces that the grains would have if not in contact.
(*b*) Steps in (*a*) become dislocations in a grain boundary.

orientation of the two crystals being shown in Fig. 13(*a*). Since these crystals are joined together in a single mass they attempt to adhere as closely as possible, and the resultant atomic arrangement is as represented schematically[18–20] in Fig. 13(*b*). It is seen that a certain number of lines of atoms coming from the top of the figure must terminate in the structure at places such as those represented by the inverted T's in the figure. At these places the atom on the bottom of the line has no neighbour beneath it, and as a result one of its electron-pair bonds cannot be properly completed. In a 3-dimensional crystal each of these atoms would be a line of atoms extending into the Figure, and each line which terminated would really be the edge of a plane.

The incomplete bonds at the edge of the extra planes act as deathnium centres and to some degree as acceptors. The result is that a germanium grain boundary produces various sorts of electrical disturbance which are undesirable in most transistor structures.

In the terminology used in discussing imperfections in crystals, the lines of atoms at the edge of the incomplete plane are called *dislocations*. Dislocations have been a subject of study in crystal physics for many years, but only in the last five years or so has clear-cut experimental observation of dislocations been possible. One of the most striking examples occurs in grain boundaries having very small angles in germanium, when it is possible to "see" the individual dislocations on the grain boundary. If the angle between the two crystal grains is very small, it is evident that the steps in Fig. 13(*a*) will be very far apart. In some cases, angles of less than 0·01° have been observed, with the consequence that the dislocations are spaced approximately 10^{-4} cm apart. This spacing is so large that it can readily be resolved with the aid of an optical microscope. The actual dislocation itself, however, has only atomic dimensions and would be much too small an object to be visible.

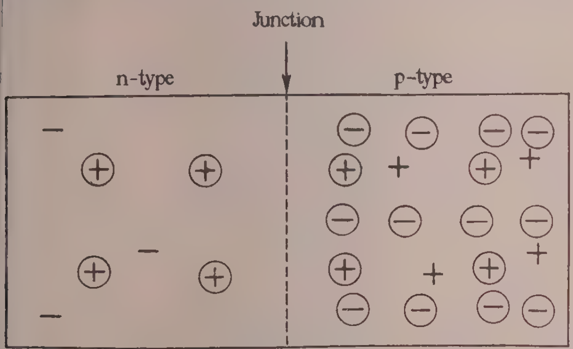
An individual dislocation can be made evident, however, by chemically etching the surface of the germanium crystal.[21,22] In the neighbourhood of the dislocation the planes of atoms are somewhat distorted, so that the electron-pair bonds are stretched in some cases and compressed in others. The energy is the sum of the elastic energy of deformation stored in the neighbourhood of the dislocation and the energy of broken bonds on the edge of the dislocation itself. This extra energy makes the germanium in the neighbourhood of the dislocation less stable and more susceptible to attack by a chemical etch than the normal germanium. As a result, if a carefully polished surface of germanium cut perpendicular to the grain boundary is etched with suitable reagents, pits tend to form at the location of the dislocations where the germanium is more readily dissolved. These pits can be seen in an optical microscope.

Fig. 14.—A row of etch pits, each corresponding to a dislocation, along a small angle-grain boundary.

Fig. 14 shows an example of such a grain boundary, the etched pits being about 10^{-4} cm apart.[21] The difference in orientation between the two grains on the sides of the grain boundary was determined with the aid of X-ray techniques, and it has been found that this difference is just what would be expected if there was one dislocation and consequently one step for each of the dots shown in the Figure. If one looks at other parts of the crystal one sees a few additional etch pits in the clear areas, but there are also large areas with no such pits; in these, the crystal is apparently ideal except possibly for chemical impurities, an occasional germanium atom which is out of place and in an interstitial position, or an occasional location from which an atom is missing. These latter possibilities are very small, and as a consequence it is probable that in single crystals like that shown in Fig. 16 not more than one atom in 10^{12} has any imperfection in its valence bonds due to dislocations or atomic disorder.

I shall next consider how the addition of chemical impurities can be carried out so as to produce compositional structures like the *p–n* junction discussed in the previous Section. One of

simplest ways of making such a junction is to utilize the …iciple of compensation discussed in Section 2. This can be …e by growing a crystal from the crystal-growing machine, …g a melt containing, for example, arsenic, so that the resulting …gle crystal is n-type. During the crystal-growing process a …et of impurity from the third column of the periodic table, … indium, is added to the melt.[23] The amount added is much …ater than the original arsenic content and the material which …sequently solidifies thus contains more acceptors than donors. … a consequence of the principle of compensation the crystal …ich forms subsequent to the addition of the indium is p-type. …e crystal obtained in this way thus has a p–n junction. This …ction is not between simple n- and p-type materials, as was … example discussed earlier, but has, as shown in Fig. 15,



…g. 15.—A p–n junction produced by crystal pulling and compensation.

…nors as well as acceptors in the p-type region. This technique … used for making junction diodes and photo-diodes which are …nsitive to light, and is also used in one of the techniques for …aking junction transistors.

Two other methods are of particular interest in forming com…sitional structures. Both of these start with uniform crystals … germanium and add the impurities without raising the tem…rature to the melting point of germanium. One of these …chniques—known as alloying—has been extensively used in …aking junction transistors, and more transistors have probably …en made by this technique than by any others; but in spite of … importance I shall discuss it only very briefly. In essence the …chnique is as follows: On the surface of the specimen of n-type …rmanium one places a small pellet of indium.[24–26] This com…nation is raised above the melting point of indium, and some … the germanium dissolves in it. The temperature is then …wered and the germanium precipitates. Since it precipitates …om solution in indium, it carries with it some indium atoms, … that the recrystallized germanium is strongly p-type. Thus …e undissolved germanium forms a p–n junction with the …crystallized germanium and the indium pellet may be used as … electrical contact to the p-type germanium.

Compensation may also be produced without dissolving and …crystallizing the germanium by making use of higher tem…ratures. If donors or acceptors are put in contact with the …rface of germanium at high temperature, they will enter the …rface and penetrate the germanium through *solid diffusion.* … such diffusion the impurity atom advances in a Brownian …otion from one atomic position to the next through the crystal. …e exact nature of the process of diffusion need not concern us …re; it is currently a subject for research and speculation. In …ssing it is worth while to point out that several mechanisms are …ssible. One is simply an interchange of adjacent atoms whereby …e impurity and adjacent germanium acquire by chance …fficient thermal energy to push past each other and change places. Another possibility is that there may be vacant sites from which atoms are missing; if one of these moves as a result of the jumping of germanium atoms to a position next to an impurity, the impurity may then jump into the vacant site and by this means move one inter-atomic distance. No matter what the mechanism, however, the observed fact is that if silicon or germanium crystals are put into contact with the vapour of donors or acceptors at sufficiently high temperatures and for sufficiently long times, then donors and acceptors enter the germanium in substantial numbers and diffuse into it obeying the laws of diffusion. If the germanium is not initially very heavily doped, n-type material may be converted to p-type or p-type to n-type by exposing it to an atmosphere of impurities of the opposite type.

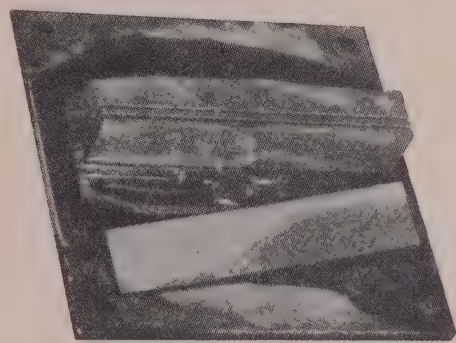


Fig. 16.—A silicon crystal sliced into plates prior to diffusion treatment.

Fig. 16 illustrates the first step in producing a p–n junction in silicon by solid diffusion.[27,28] This Figure represents an ingot of n-type silicon mounted on a piece of glass and sliced into plates with a diamond saw. Fig. 17 shows one of these plates

Fig. 17.—A silicon plate sealed in a quartz tube for diffusion treatment.

sealed in a quartz tube with a small quantity of boron-trichloride. The quartz tube is heated to about 1 000° C for about 10 min. Some of the boron-trichloride decomposes and free boron is formed on the silicon surface. This boron diffuses inward, producing a very heavily doped and very thin surface layer of p-type material. Thus a p–n junction is formed just below the surface of the silicon entirely covering the silicon slab.

Fig. 18 illustrates a method of making electrical connection to such a p–n junction; the p-type layer is cut away on one surface of the plate, electrical connection being made to this and also to the p-type layer. Although the p-type layer is extremely thin, its conductivity is so high that contacts at one place can be used to apply bias over the entire layer.

Such diffused junctions on silicon have outstanding properties as semi-conductor junctions. They are extremely active photo-electrically and are used in the solar battery. They also have outstanding properties as electrical rectifiers. I shall describe some of these features in the following Section.

(5) THE DIFFUSED SILICON JUNCTION

The fact that the diffused silicon junctions are extremely near the surface makes them very effective in converting light into electrical energy.[29,30] Another reason is that the potential-

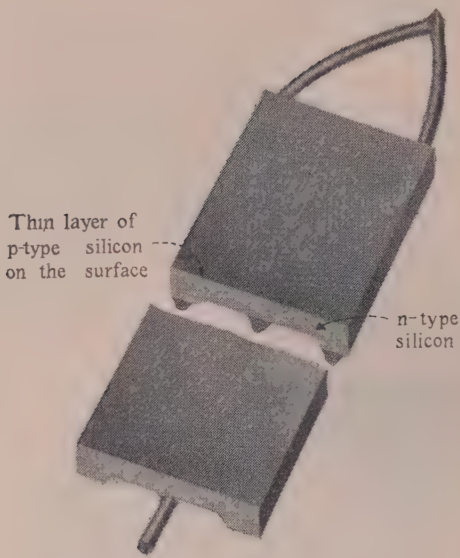


Fig. 18.—A method of making connection to a diffused *p–n* junction on silicon.

energy hill in silicon under conditions of thermal equilibrium is higher than in germanium. Let us consider first the role of the thinness of the junction, or rather the thinness of the *p*-layer on the surface. Most of the light falling on the surface and having sufficient energy to produce a hole-electron pair is absorbed within less than 10^{-3} cm of the surface, and thus most of the hole-electron pairs produced by light will be within about 10^{-4} cm of the surface. This distance is so small that the minority carrier in the hole-electron pair will have a very good chance of reaching the *p–n* junction before recombining at a deathnium centre with a majority carrier. Once it reaches the junction it will be drawn across it by the electric field that exists under conditions of thermal equilibrium. The net result will be to transfer an electron from the *p*-type material to the *n*-type material if the photon is absorbed in the *p*-type material, or to transfer a hole from the *n*-type material to the *p*-type material if the photon is absorbed in the *n*-type material. In either case the net result is the same, and the hole-electron pair is separated by the field of the junction with the electron ending in the *n*-side and the hole in the *p*-side. As a result of this separation the photocurrent gives the junction a forward bias, which will build up until it produces a forward current equal and opposite to the photocurrent.

Since the potential hill is higher in silicon than in germanium, it is evident that a larger forward bias can be built up in silicon than germanium. Experimentally it is found that sunlight shining on a silicon junction will develop a forward potential of approximately 0·5 volt before the forward current balances the photocurrent produced by the light. If the junction is short-circuited through an ammeter, it is found that sunlight will produce a current of approximately 60 mA/cm². If such a junction is matched into an optimum load, it produces a potential of about 0·3 volt and a power output of approximately 100 mW/cm². Thus a square metre of such a silicon *p–n* junction exposed to bright sunlight will produce a power output of 100 watts. The total amount of solar energy falling upon a square metre is approximately 1 kW, so that the diffused silicon junction converts solar energy to usable electrical energy with an absolute efficiency of 10%.

Efficiencies as high as 12% have been obtained with experimental junctions, and such efficiencies are approximately ten times higher than any previously obtained in conversion of solar energy to electrical power. They are, of course, enormously higher than indirect processes involving biological effects, such as the growing of crops with the aid of sunlight which are subsequently used as fuel in engines. It is not yet clear whether solar power will be a useful form of power at high power levels, for much depends upon the competition produced by atomic energy and fossil fuels. However, for certain limited power applications it appears probable that the solar battery will be a truly useful device.

One of the first practical experimental applications is planned for the summer of 1955; an array of silicon discs composing a solar battery will be mounted on a telephone pole, and the solar battery will, in turn, be connected to a storage battery used for powering a transistor amplifier. The arrangement is contemplated for use on rural lines where, for purposes of economy, carrier circuits are employed. It would not be practical to use the solar battery to power an amplifier using vacuum tubes, since their lower efficiency leads to a power consumption which it would be not practical to supply from the solar battery. The apparatus is designed so that the storage battery will remain charged if as little as 15 hours of sunlight a week is available.

The thinness of the layers possible in the diffused silicon junction, together with the high energy gap, make these devices outstanding in the role of rectifiers at high power levels. For example, two discs of silicon about ½ in in diameter and used in series will make a half-wave rectifier which will readily give 300 watts of half-wave rectified power working directly from a 230-volt line. This operation is a conservative one for such units, since each will operate at a reverse voltage of 200 volts. Furthermore, even at a temperature as high as 200° C the reverse resistance of these units remains so great that power losses in them during the reverse-voltage part of the cycle is not excessive. It is thus possible to produce power rectifying units of substantially smaller size and lower heat-dissipation requirements than has hitherto been possible.

(6) THE JUNCTION TRANSISTOR[12,31,32]

The simplest transistor to describe in terms of compositional structure is the junction transistor. In its general behaviour as an amplifying device a junction transistor shows great similarities to a vacuum-tube triode, or thermionic valve. Fig. 19 shows an *n–p–n* junction transistor in an amplifying circuit, the transistor being in the form of a sandwich with a layer of *p*-type germanium interposed between two layers of *n*-type germanium. Non-rectifying electrical contacts are made to the three layers.

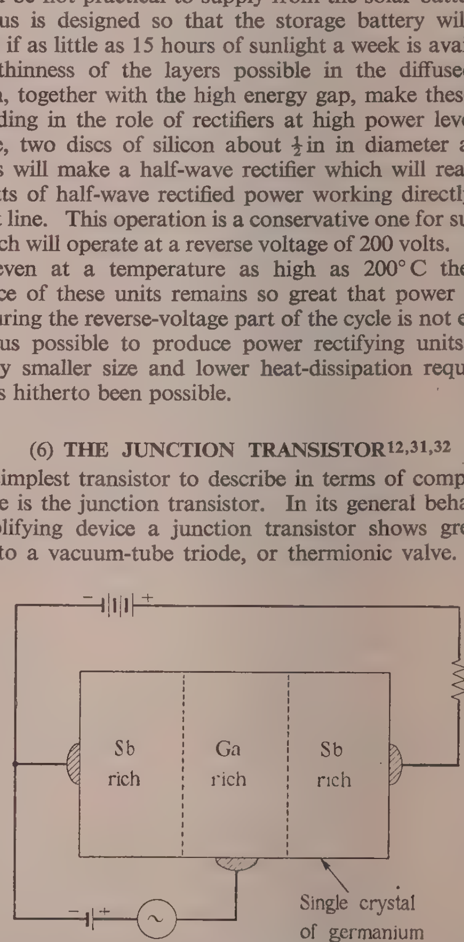


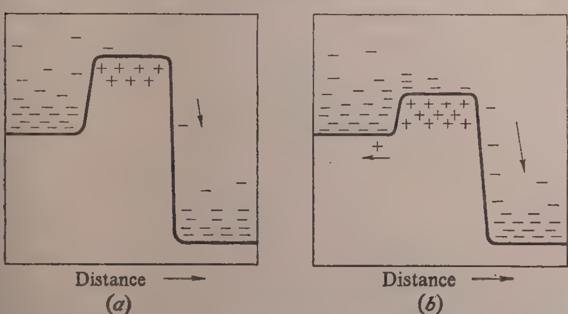
Fig. 19.—The structure of a junction transistor and the bias supply for its operation in an amplifying circuit.

der operating conditions the n-type region on the right, own as the *collector*, is biased positive so as to become tractive to electrons. As a result, a reverse bias appears between nd the middle region, known as the *base*. The current flowing ross this reverse-biased junction can be controlled by an input nal applied between the base layer and the n-type region to the t, known as the *emitter*. As I shall describe below in more tail, the bias across the emitter junction controls the electron w into the base region. In effect, the emitter junction acts like e region between the cathode and the grid in the vacuum tube. lectrons which enter the base region have a high probability of ffusing to the collector junction, and thus the flow of electrons om emitter to collector can be varied by varying the potential ross the emitter-base junction. The action is very similar to at controlling the flow of electrons from the cathode to the node in a thermionic triode.

Junction transistors can be fabricated in a variety of ways. he compositional structure can be produced in a crystal-growing achine by techniques like those used for making simple p–n nctions. As the crystal is grown from a melt containing ntimony, a pellet containing indium is dropped into the melt, nd a second pellet containing antimony is dropped in a few conds later. The portion of the crystal which grows between e dropping of the two pellets is rich in indium and is conse-uently p-type. The second pellet over-compensates the effect f the added indium and the subsequent material is again n-type. rom such a single crystal small rods may be cut and contacts nade. There are a number of technical processes involved in roceeding from the point of growing the crystal to the pro-uction of a packaged, stable transistor. I shall not endeavour discuss these in the lecture.

An alternative technique for producing the compositional ructure starts with a thin plate of germanium which subse-uently plays the role of the p-type region. A pellet of metal ontaining a donor is placed on this plate. The plate and pellet re then raised to such a temperature that the metal melts and issolves a small amount of germanium. When the metal and ermanium are subsequently cooled, the germanium precipitates om the metal and grows back onto the crystal structure of the ase material. This regrown germanium carries with it some of e donors contained in the molten metal and thus grows an -type region. In the fabrication of a transistor, pellets are laced on both sides of a thin plate and allowances are made for e degree to which they dissolve germanium on the two sides. his process has been used in the production of a large fraction, not the majority, of transistors made to date.

From the point of view of an electron the situation in an perating transistor is as represented in Fig. 20. This diagram hows the variation in potential energy for an electron along a ne going from emitter to collector in a transistor biased like



g. 20.—The potential energy of an electron in an n–p–n junction transistor for two values of bias across the emitter junction.

In (*b*) the forward bias is greater than in (*a*).

that shown in Fig. 19. The reverse bias at the collector junction produces a large drop in potential on the right-hand side. The varying bias across the emitter junction changes the height of the hill and thus varies the diffusion current of electrons into the base. The base layer is very thin and contains very little death-nium. Consequently, the probability is much higher for an electron to diffuse through the base layer and arrive at the collector junction than for it to combine with a hole through a deathnium centre in the base layer. In a well-built junction transistor, in fact, the electron flow through the base region proceeds so efficiently that the electron current flowing through the base layer to the collector may exceed the current combining in the base layer by a factor of 100 or more. This means that the input currents flowing to the base layer through the base lead may control the output currents 100 times larger flowing to the collector region. Although this situation is not nearly as ideal as in a vacuum triode, in which the grid current is still smaller compared with the anode current, it permits the junction transistor to operate as a highly efficient amplifier.

So far as voltage requirements are concerned, the junction transistor is far superior to any vacuum tube, because a junction transistor can be brought fully into its operational range with voltages as small as 50–100 mV. Such voltages make the potential hill at the collector junction shown on Fig. 24 several times the thermal voltage ($kT/e = 25$ mV). As a consequence, any electron reaching the collector junction is sure to slide down the hill and has a negligible chance of returning.

The junction transistor has almost ideal pentode charac-teristics in terms of collector saturation. This also follows from the same considerations which make it operate well at low collector voltages. In fact, in some junction transistors there is a negligible variation in current over a voltage range of a hundredfold or more, say from 100 mV to 10 volts on the collector.

The junction transistor described in Figs. 19 and 20 is only one of a large family of transistors. The p–n–i–p and n–p–i–n transistors are variations of this form especially designed for low capacitance between the collector and the base. The so-called junction tetrode is a special form of junction transistor in which the current flow is controlled so as to occur only over a small region of the base. Junction transistors of these types have been used in oscillating circuits at frequencies as high as 1 000 Mc/s.

In closing this Section I should like to point out that the junction transistor operates more flexibly in terms of power than do vacuum tubes. This is of great importance, since it means that junction transistors may be used efficiently in some cases where the power level to be amplified is very much smaller than the heater current required in the most efficient of vacuum tubes. The possibility of operating at very low power is due to the low voltages necessary to operate a junction transistor, and because of its small size and the high degree of purity possible, the current due to generation by deathnium centres can be made smaller than a microampere. It is thus possible to make a junction transistor which can be put into an amplifying con-dition with a power input substantially less than a microwatt. The same transistor, however, can operate at voltages as high as 10 volts and currents of the order of 10 mA. Thus it can cover a power range effectively of 100 000. This fact indicates that there will probably be a much smaller diversity of transistor types than there are of vacuum-tube types.

(7) COMMERCIAL APPLICATIONS

There have been two large commercial applications of transistors in the United States, namely the hearing aid and the portable radio. There are rather more than a million hearing

aids in use in the United States; approximately one third are now designed for transistors and use rather more than a million transistors. The chief manufacturer of these transistors employs the alloy method of manufacture, and the next largest manufacturer the double-doping method. The transistor has an advantage over the vacuum tube in hearing aids because no filament power is necessary. Furthermore, the supply required for the collector voltage may be of considerably lower voltage than the anode supply for vacuum-tube hearing aids. As a result, battery replacement occurs much less frequently with the transistor hearing aid, and, moreover, the batteries are less expensive. It has been estimated that when the hearing aids become fully adapted for transistors, the hearing-aid users of the United States will benefit from an annual saving in battery cost of about $50 million a year. Experience to date indicates that transistors are reliable in hearing aids and the replacement rate of transistors appears, in fact, to be lower than that of vacuum tubes.

At least two portable radio sets utilizing transistors have now appeared. The first of these is the smallest commercially produced model; it uses four transistors and a 22½-volt power supply, and will operate for approximately 30 hours on one battery. The second model uses a much larger loudspeaker and is competitive in sound volume with vacuum-tube radio sets; its four 1½-volt flashlight batteries will keep it in operation for 500 hours. It thus has the advantage over any vacuum-tube portable radio set that batteries need replacement much less frequently and are of a type almost universally available.

Some idea of the potential growth in the transistor field may be obtained by considering Figs. 21 and 22. Fig. 21 shows the

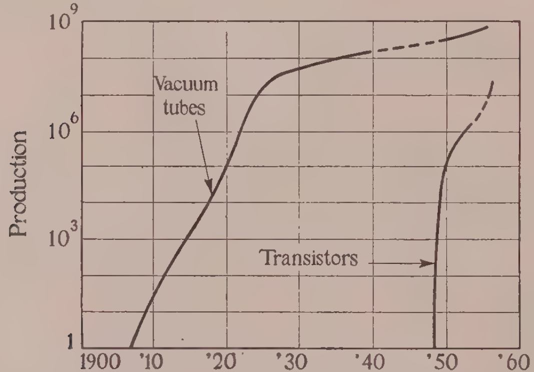


Fig. 21.—The annual rate of production of transistors and vacuum tubes.

annual production of vacuum tubes from the date of invention of the de Forest audion; in that year one vacuum tube was made. The data for the years up to about 1915 were difficult to obtain from the literature, and my source was Dr. de Forest himself. The transistor production is now more than a million and less than 10 million a year, and thus corresponds to vacuum-tube production at about 1920.

The cost of a vacuum triode has declined steadily year by year. At the same time, industrial wages have continued to increase. In order to reflect increased efficiency of production of vacuum tubes Fig. 22 shows their cost, not in dollars, but in terms of the cost of a man-hour of industrial labour. It is seen that the cost of a transistor in these terms is dropping very rapidly. The figure used is the retail price of the cheapest junction transistor. We may expect transistor prices to equal tube prices by about 1956 or 1957, after which there should be an explosive growth in transistor production, the mechanism being somewhat as

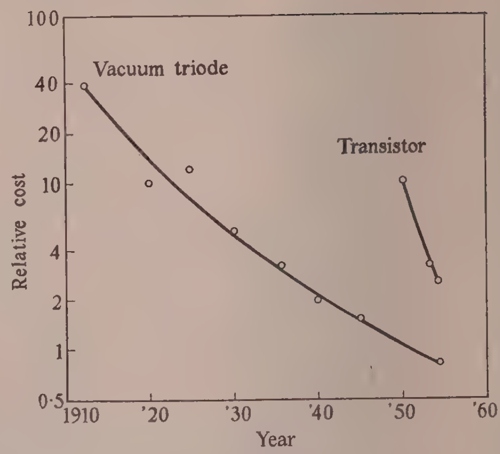


Fig. 22.—The prices of a vacuum triode and a transistor as a function of time.

follows: when transistors become cheaper than tubes they will take over many of the regular vacuum-tube applications, under which conditions the production will rise abruptly. We again note that transistor production now corresponds in volume to vacuum-tube production in about 1920, and that there has been a drop in price by a factor of 10 in vacuum-tube costs between 1920 and the present day. Thus, when transistor production increases greatly it seems reasonable to suppose that transistor prices will drop rapidly, possibly by a factor of 10 when their production equals that of vacuum tubes.

Thus it is to be expected that there will be a violently rapid growth of transistor activity starting in about 1956 when the price of a transistor is expected to fall below that of a vacuum tube.

(8) RECENT DEVELOPMENTS IN TRANSISTOR PHYSICS

The understanding of the fundamental physics of semiconductors has advanced at a rate quite comparable to that of the applications. In order to describe this progress with real exactness, I should have to discuss a number of new concepts in considerably more detail than was necessary for the five basic imperfections. For the purposes of the lecture, however, exactness in regard to theoretical detail is less important than accuracy of impression about the diversity of effects and the extent of detailed knowledge. This detailed knowledge is generally quantitative in form and in many cases is concerned with the interactions of imperfections. In particular, several features of the interaction of holes and electrons with donors, acceptors and deathnium centres have been measured. In order to discuss the techniques for making these measurements and interpreting them, it would be necessary to go into more detail than is practical within the scope of the lecture; however, the extent of the activity can readily be appreciated from the large number of results that have been obtained. In order to illustrate this conclusion, a number of examples will be quoted.

Some of the most important interactions that have been studied are illustrated in Fig. 23; they are those of an electron with donor and acceptor centres. Similar effects occur for holes, and electrons may also interact with other imperfections not represented in Fig. 23. In general, two broad classes of behaviour are possible for an electron in the presence of donors. The electron may either be free, so that it does not remain attached to any particular donor, or it may be bound, when its behaviour is rather similar to that of an electron in an atom. It may exist in its most tightly bound state represented in the upper part of

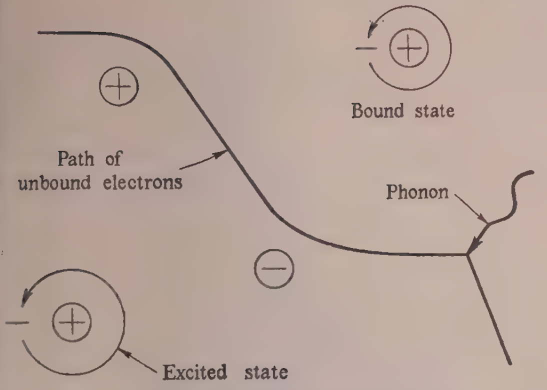


Fig. 23.—Some interactions of importance in transistor physics.

: Figure or it may be bound in an excited state having some-at higher energy. In the case of atoms, transitions from the vest state to excited state are accompanied by the absorption light, and transitions from excited states to the lowest state th the emission of radiation. Similar effects occur for electrons und to donors in semi-conductors, and it is possible by this eans to determine the energy difference between the states volved.

If the electron is free and moving through the crystal, its th will be deflected as a result of the acceleration given it by ectric forces acting between it and a positively charged donor. may also be acted upon by an acceptor, and in this case the rce will be a repulsion. Fig. 23 illustrates the way in which ese interactions may deflect the path of an electron. Such teractions interfere with the motion of the electron through e crystal and give rise to electrical resistance.

Electrical resistance will also arise from the interaction of the ectron with thermal agitation of the atoms of the crystal. nese thermal agitations can best be described in terms of sound aves moving through the crystal. These waves are quantized, st as are waves of electromagnetic energy. Each quantum of sound wave carries a certain amount of energy, the energy ing equal to Planck's constant times the frequency of vibra-on in the wave, just as it is for light. The name given to one antum of thermal-vibration energy in a crystal is the *phonon*. phonon may collide with an electron, impart its energy to the ectron and thus be absorbed. As a result, the path of the ectron is deflected. Alternatively, an electron may generate a onon and change its direction of motion. Both these processes ntribute to the electrical resistance, since such collisions tend destroy any coherent motion produced on the electrons by an ectric field. Such an interaction is also illustrated schematically Fig. 23.

One of the most important ways of studying binding energies tween electrons and donors is illustrated in Fig. 24, which presents a specimen of germanium containing five donors and o acceptors. The distribution of holes and electrons is presented for conditions of thermal equilibrium at several ferent temperatures. The case shown is representative of a pical sample of germanium.

At room temperature the situation is that described in earlier ctions of the lecture. A condition of electrical neutrality is oduced, and under this, since the specimen is n-type, the nsity of electrons is much greater than that of holes. The case own represents only a small portion of the crystal, and the le density is so low that no holes are observed in the area own. At very low temperatures the thermal agitation is sufficient to shake the electrons off the donors, and thus all

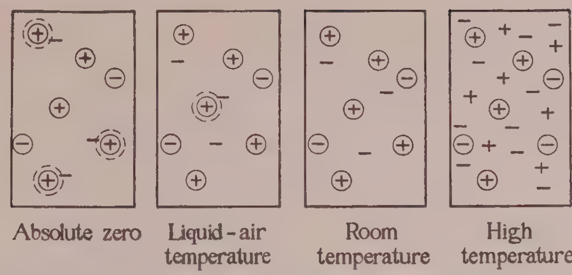


Fig. 24.—The variation in carrier density in a partially compensated n-type specimen as a function of temperature.

the electrons are bound to donors in the lower state. This is the situation corresponding to absolute zero. As the temperature is raised, thermal agitation occasionally frees an electron, which moves through the crystal and then again becomes bound to a donor. Under thermal-equilibrium conditions a certain fraction of the electrons will be bound and a certain fraction free. At the temperature of liquid air approximately two thirds of the electrons are free. At high temperatures the carrier densities required by the mass-action law are so much greater than those required by doping that the effect of impurities is no longer important and the behaviour is the same as it would be in a pure specimen; intrinsic conductivity then occurs.

The law governing the fraction of electrons freed from donors at a given temperature involves a Boltzmann factor in a manner similar to that for the product of hole and electron densities. The energy in the Boltzmann factor, however, is not that required to create a hole-electron pair but that required to free an electron in the lowest bound state from a donor, i.e. the energy necessary to ionize one of the neutral donors in Fig. 24. Owing to this Boltzmann-factor dependence, the electron density varies approximately as $\varepsilon^{-W_B/kT}$, where W_B is the binding or ionization energy of a donor. As a consequence, if the logarithm of the electron density is plotted as a function of $1/T$, a straight line having a slope of $-W_B/k$ should be obtained.

Some experimental data[33] taken from two germanium specimens are shown in Fig. 25; the density of electrons for the

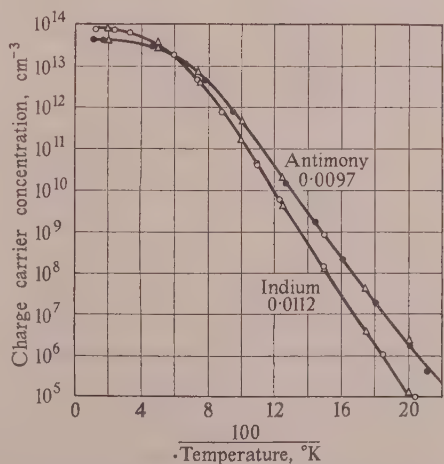


Fig. 25.—The method of analysing carrier density versus temperature so as to determine binding energies.

antimony sample and of holes for the indium sample are shown plotted on a logarithmic scale for the ordinate, with the reciprocal of temperature as abscissa. Both experimental and theoretical

points are shown. These theoretical points are based on the Boltzmann factor corresponding to binding energies of 0·0097 eV for antimony donors and 0·0112 eV for indium acceptors. Other corrections involving slowly varying factors have been included in the calculations. It is seen that at the lower temperatures the data are well represented by straight lines, which corresponds to the prediction made above. It should be noted that the data cover a really enormous range: the hole density in the indium sample varies by a factor of 10^9 between room temperature and 5° K.

In order to deduce the carrier densities in Fig. 24 from experimental measurements it is necessary to measure, not only the electrical conductivity of the specimens, but also another quantity known as the Hall coefficient. The Hall coefficient is measured by applying a magnetic field transverse to the direction of current flow in the specimen. This magnetic field tends to deflect the carriers to one side or the other depending upon their charge, the direction of drift due to the electric field and the direction of the magnetic field. As a result of this tendency to drift sideways, charges build up on the surfaces of the specimen and set up a compensating transverse electric field, which can be measured by placing electrodes on opposite sides of the specimen. By means of the theory of the Hall effect it is possible to deduce values for both the carrier concentration and the mobility of the carriers from measurements of the Hall coefficient and the conductivity of the specimen. Some unresolved theoretical points remain in the interpretation of such data, but these introduce uncertainties which are small compared to the effects observed. As a result it is now believed that the binding energies deduced in this way are accurate to within a few per cent.

A number of acceptors and donors have been measured in both silicon and germanium. The thermal ionization or binding energies[34] for these are shown in Table 3. Also, as seen in the

Table 3

THERMAL IONIZATION ENERGIES

	Solute element	Ionization energy	
		Germanium	Silicon
		eV	eV
Group III acceptors	Boron	0·0104	0·045
	Aluminium	0·0102	0·057
	Gallium	0·0108	0·065
	Indium	0·0112	0·160
Group V donors	Phosphorus	0·0120	0·044
	Arsenic	0·0127	0·049
	Antimony	0·0096	0·039
	Bismuth	—	0·067
	Lithium	0·0093	0·033

Table, four elements of valence three act as acceptors in both silicon and germanium, and four elements of valence five as donors. Lithium is a special case; it diffuses very readily in both silicon and germanium and it is conjectured that the lithium atom does not substitute itself for atoms of the lattice, but occupies interstitial positions between the normally occupied sites. Lithium is a monovalent element having one loosely bound electron; this electron apparently becomes free, leaving behind a positively charged lithium ion. Thus lithium acts as a donor.

The behaviour of other elements from the periodic table is somewhat more complicated. The numbers describing the characteristics of several elements[35] are shown in Table 4. In

Table 4

CONSTANTS DESCRIBING VARIOUS IMPURITIES IN GERMANIUM AND SILICON

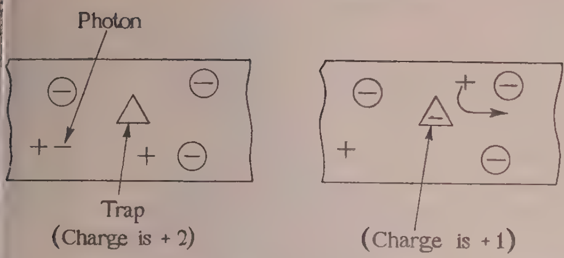
	Solute element	Ionization energies		Capture cross-sections	
		$(W_c - W)$	$(W - W_v)$	Holes	Electrons
		eV	eV	Å²	Å²
Germanium	Group III		~0·01	<0·1	<0·1
	Group V	~0·01		<0·1	<0·1
	Nickel	0·30	0·22	>40	0·8
	Copper		0·040 0·25	1·0	0·1
	Gold	0·20	0·15		
	Iron	0·27	0·34		
	Cobalt	0·31	0·25		
	Zinc		0·029		
	Platinum		0·040		
Silicon	Group III		0·04–0·16		
	Group V	0·03–0·07			
	Gold	0·3	0·39		
	Zinc		(0·092)? (0·3)?		
	Heat treatment	0·033 0·133 0·3			

some cases several energies are shown; it is not certain whether these correspond to several states of ionization of one kind of centre or to several kinds of centres produced by one impurity.

In germanium, nickel is an active form of deathnium and appears to be in the acceptor class and to bind a hole with an energy[35] of 0·25 eV. The numbers describing the role of nickel as deathnium are also shown in Table 4. When the nickel atom is negatively charged at room temperature, its ability to capture a hole is described by a *cross-section* which is greater than 40 Å², the significance of which is as follows: When the nickel is negatively charged, on the average, if a hole passes within a certain distance of it, the hole will become bound. This distance is such that, if a circle having this distance for its radius is drawn, the area of this circle is approximately 40 Å². An Ångström unit is 10^{-8} cm and is approximately the diameter of a hydrogen atom. When the nickel atom has captured a hole, however, so that it is electrically neutral, its ability to capture an electron is described by the number 0·8 Å². Thus, nickel is a more effective form of deathnium in *n*-type samples than in *p*-type samples. This conclusion follows from the fact that in *n*-type samples the nickel centres will, in general, be negatively charged and thus will be, so to speak, cocked traps ready to capture holes; on the other hand in *p*-type specimens the nickel centres will, in general, be neutral and will not be so effective in capturing electrons.

The amount of work which has been carried out in determining energies and characteristics of various chemical impurities can be appreciated at once from Table 4. It is also to be noted that the story is far from complete and that certain energies have been determined for centres in silicon that are produced by heat treatment. When the Table was prepared the identity of these impurities or imperfections had not been established.

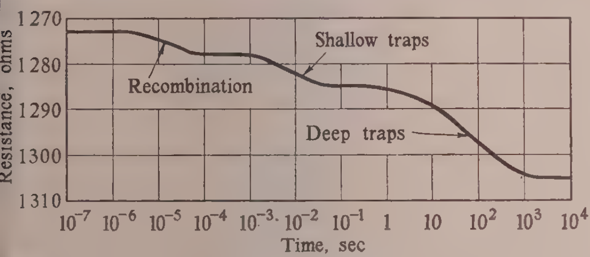
There are other centres prominent in silicon that are characterized by a process known as trapping.[36] One hypothetical mode for a trap is illustrated in Fig. 26, which shows a sample of *p*-type silicon containing initially three acceptors, one hole and an imperfection which is assumed to have a charge of +2. A photon is absorbed in this specimen, producing a hole-electron pair. The positive charge on the trap results in the prompt capture of the electron. The trap, however, has a net charge



g. 26.—The conjectural behaviour of a trapping centre in silicon.

+1 after this trapping process. Thus it is not an effective ıthnium centre, since after capturing the electron it still has ositive charge and tends to repel holes. So long as the electron nains in the trap, the conductivity of the specimen will be reased over its thermal-equilibrium value, since an extra hole ll be present in order to make up for the change in charge on trap due to the trapped electron. Thus the photo-conductivity oduced by the light will persist so long as the electron remains the trap. Traps in silicon make their presence felt by causing ry long time delays in the photo-conductive process.

Some data[36] pertinent to this problem are represented in g. 27. Here the resistance of a specimen is shown as a function



g. 27.—The effect of trapping upon the decay of photoconductivity in a *p*-type silicon specimen.

time immediately after illumination by a pulse of light less an 10^{-7} sec in duration. It is seen that immediately after this sh of light the resistance is lower and the conductance higher an it is in the thermal-equilibrium condition; the thermal-uilibrium value is restored after about 10^4 sec. During the st 10^{-7} sec, electron-hole pairs are produced by light and a mber of the electrons are trapped. This first stage occurs too pidly to be resolved on the oscillograph. After the light is rned off there is little change in the conductivity for the first icrosecond, but after 100 microsec there is a substantial drop conductivity, owing to recombination of electrons which were t trapped during the first fraction of a microsecond. Very tle change in photo-conductivity then takes place until about millisec has elapsed; after this a substantial number of electrons hich have been trapped at centres, known as shallow traps, are ed by thermal agitation and recombined with electrons. An ctron has approximately a 50% chance of boiling out of any e of these traps in a time of about 10^{-2} sec. As a result, no preciable emptying of the shallow traps occurs during the first $^{-3}$ sec and the shallow traps are substantially exhausted after $^{-1}$ sec have elapsed.

An additional group of traps, indicated as deep traps, make eir presence felt after about a second. It takes nearly an hour fore all of these traps have ejected their electrons and the ecimen returns to its original condition. Although the binding ergies of electrons in the traps and certain other of their ributes have already been measured, at the time when Fig. 27 was prepared the physical or chemical nature of these traps was uncertain.

The last topic I shall discuss has to do with what in transistor physics is the most basic interaction of all—the interaction of the electron with the perfect lattice. The motion of an electron is governed by the Schroedinger wave equation, which incorporates the basic quantum mechanical laws that apply to microscopic particles. In moving through the crystal the electron is subjected to large forces of attraction by the atomic nuclei and repulsion by the other electrons. These forces produce interference effects on the electron waves, and as a result the electron waves are refracted and move through the crystal at a velocity which differs from the free-space value. To a first approximation it is supposed that this refractive effect can be described by a sort of refractive index. The refractive index can be introduced in the theory by saying that in the crystal an electron accelerates at a different rate in response to an electric field than it would in free space. This difference can be described by saying that the electron acts as if it had a greater or lesser inertial mass than it would have in a vacuum.

Recent developments in theory and experiment on silicon and germanium have shown that it is too crude an approximation to assume that the interaction of the electron with the perfect crystal can be described by a single mass number. Instead, one must reach the conclusion that an electron is described by several different masses, i.e. it accelerates as though it had one mass for an electric field in one direction and a different mass for an electric field in another direction. Furthermore, the electron may shift from one mode of motion to another, and in doing so can change its effective masses. In germanium there are four possible modes of motion. In one of these the electron has a large mass for acceleration along a certain crystal axis and a small mass for accelerations at right-angles to this. So long as the electron stays in this mode of motion its behaviour in electric and magnetic fields can be predicted from these two masses. However, when the electron interacts with a phonon it may take a transition from one of these modes of motion to another one in which its mass numbers are different in different directions.

The theoretical possibility had been appreciated for many years that electrons moving in crystals might have different modes of motion and in each of these modes different characteristic masses for accelerations in different directions, but clear-cut evidence for these effects including measurements of the mass numbers have developed only recently. These new developments have been made possible by the availability of good single crystals of silicon and germanium. The new techniques involve the use of very-high-frequency electric fields, low temperatures and high magnetic fields.[37]

The way in which the mass numbers may be measured is illustrated schematically in Fig. 28, which shows four electrons moving at very low temperature in a specimen of *n*-type germanium. It is assumed that a magnetic field is applied to the specimen perpendicular to the plane of the Figure. For the particular orientation chosen in one mode of motion the electron is deflected sideways by the magnetic field acting in all cases upon the small transverse mass. The other three modes of motion have masses which are anisotropic so far as the magnetic field is concerned. In these cases the projection of the electron orbit for each mode is an ellipse, and the frequency of the periodic motion is slower than for the circular case.

The motions shown in Fig. 28 are analogous to the circular motions of ions in a cyclotron. In a cyclotron an ion may absorb energy from the applied electric field if its frequency of motion in the magnetic field is in resonance with the frequency of the electric field. The same thing may occur in a semiconductor and the phenomena is by analogy referred to as

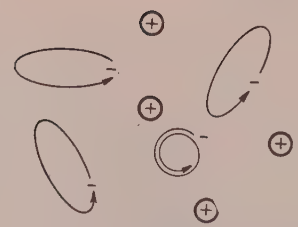


Fig. 28.—The orbits in a magnetic field for the four modes of motion in a germanium crystal.

Magnetic field along [111] body diagonal of cubic cell.

cyclotron resonance. Prior to the carrying out of experiments it was realized on the basis of theoretical reasoning that in specimens of germanium or silicon it should be possible to achieve experimental conditions in which electrons would traverse several radians of a cyclotron orbit before being scattered by impurity interactions or interactions with phonons. Under these conditions, the motion would have a relatively "high Q-factor" and the resonance would be relatively sharp. These resonances occur in the microwave range and can be observed by measuring the absorption of a specimen of germanium in a waveguide cavity.[38] The measurement may be made by maintaining a fixed frequency of electromagnetic field and varying the magnetic field so as to bring one mode of motion after another into resonance with the electromagnetic field.

An example of this effect is shown in Fig. 29. Here the

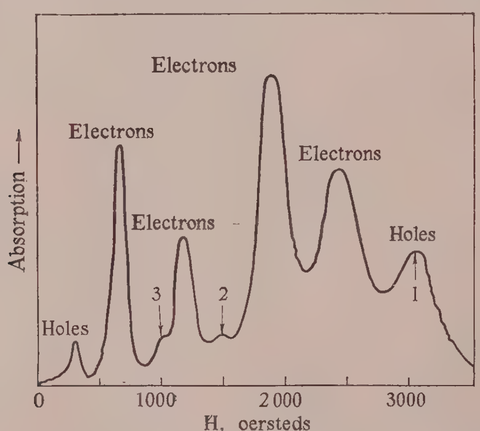


Fig. 29.—Absorption due to cyclotron resonance in a germanium specimen.

absorption due to the germanium is plotted as the ordinate and the magnetic field as the abscissa. In the specimen concerned, both holes and electrons were present. It should be noted in particular that for this orientation of magnetic field each mode of electron motion has a different effective mass, so that four resonances are observed. It should also be noted that two resonances are observed for holes, one at a very low magnetic field corresponding to a very small mass and one at a high field corresponding to the largest mass for holes.

By making measurements like that represented in Fig. 29 and by varying the direction of the magnetic field it is possible to obtain data which fit into a consistent picture from which mass numbers may be deduced for the various modes of motion. In some cases it is found that a description in terms of a longitudinal mass and transverse masses is inadequate. The motion in such cases is not elliptical, but more in the form of a rosette. Evidence for this is found also in Fig. 29 for holes. It is noted that there are two small peaks marked 2 and 3 which occur at one-half and one-third of the magnetic field for the fundamental resonance for holes. These resonances correspond to harmonics of the basic motion of the holes and arise from the fact that the orbit is not elliptical but has components of motion at higher harmonics.

It is evident that the technique of cyclotron resonance has introduced a sort of spectroscopy for the fundamental interactions of electrons and holes with a perfect crystal. It is being extended at the present time to other semi-conductors, and from it there will result a more complete knowledge of these basic interactions than has hitherto been possible for the motions of electrons or holes in crystals.

An indication of the extensiveness of the work carried out at present may be seen from Table 5, which shows that many

Table 5

EFFECTIVE MASS NUMBERS

Electrons	Germanium		Silicon	
	4 or 8 ellipsoids		3 or 6 ellipsoids	
M_l	$1 \cdot 57 m_0$		$0 \cdot 98 m_0$	
M_t	$0 \cdot 082 m_0$		$0 \cdot 19 m_0$	
M^*	$0 \cdot 12 m_0$		$0 \cdot 26 m_0$	
Holes: $W = -\frac{h^2}{2m_0}\{Ak^2 \pm \sqrt{[Bk^4 + C(k_x^2 k_y^2 + k_y^2 k_z^2 + k_y^2 k_z^2)]}\}$				
A	13·0		4·1	
B	8·7		1·4	
C	11·4		3·7	
Average mass	Mobility	Density of states	Mobility	Density of states
$\overline{M}_h$	$0 \cdot 31 m_0$	$0 \cdot 36 m_0$	$0 \cdot 49 m_0$	$0 \cdot 53 m_0$
$\overline{M}_l$	$0 \cdot 044 m_0$	$0 \cdot 044 m_0$	$0 \cdot 16 m_0$	$0 \cdot 17 m_0$
$\overline{M}^*$	$0 \cdot 25 m_0$	$0 \cdot 37 m_0$	$0 \cdot 38 m_0$	$0 \cdot 59 m_0$

mass numbers have been determined[39–42] for both electrons and holes in germanium and silicon. The modes of motion fall into four groups for germanium and three for silicon. The modes of motion for holes are somewhat more complicated, and for an understanding of the details reference should be made to the literature. In all cases in the Table the quantity m_0 stands for the mass of an electron in free space. It is seen that electrons are accelerated in crystals in some cases approximately 1[illegible] times more readily than they would in free space and in other cases considerably less readily.

(9) CONCLUSIONS

It has been my endeavour in the lecture to show that transistor electronics has had and will have many useful applications in electronic technology. I have also tried to show the interrelationship of the fundamental physics and chemistry to this technology and to suggest that in large measure the science and the application have advanced hand in hand.

While preparing this lecture, I tried to find a suitable note on which to close. In the course of doing this I came across the following quotation:

> It may be appropriate to speculate at this point about the future of transistor electronics. Those who have worked intensively in the field share the author's feeling of great optimism regarding the ultimate potentialities. It appears to most of the workers that an area has been opened up comparable to the entire area of vacuum and gas-discharge electronics. Already several transistor structures have been developed and many others have been explored to the extent of demonstrating their ultimate practicality, and still other ideas have been produced which have yet to be subjected to adequate

xperimental tests. It seems likely that many inventions unforeseen present will be made based on the principles of carrier injection, he field effect, the Suhl effect, and the properties of rectifying unctions. It is quite probable that other new physical principles ill also be utilized to practical ends as the art develops.

This paragraph was written in 1950 and used as a conclusion the transistor-physics section of my book on "Electrons and les in Semiconductors." I believe that, while its predictions ve been realized within the last five years, it is equally applicable the next decade as well.

(10) REFERENCES

(1) SHOCKLEY, W.: "Electrons and Holes in Semi-Conductors" (Van Nostrand, New York, 1950).
(2) LARK-HOROVITZ, K.: "The New Electronics" ("The Present State of Physics": American Association for the Advancement of Science, Washington, 1954).
(3) HEROLD, E. W.: *Journal of the Franklin Institute*, 1955, **259**, p. 87.
(4) SPENKE, E.: "Elektronische Halbleiter" (Springer, Berlin, 1955).
(5) "Proceedings of the International Conference on Semi-Conductors," *Physica*, 1954, **20**, p. 801.
(6) "Transistor Issue," *Proceedings of the Institute of Radio Engineers*, 1952, **40**, p. 1283.
(7) SCAFF, J. H., THEURER, H. C., and SCHUMACHER, E. E.: *Transactions of the American Intsitute of Mining Engineers*, 1949, **185**, p. 383.
(8) BARDEEN, J., and BRATTAIN, W. H.: *Physical Review*, 1948, **74**, p. 230.
(9) HAYNES, J. R., and SHOCKLEY, W.: *ibid.*, 1949, **75**, p. 691.
(10) PRINCE, M. B.: *ibid.*, 1953, **92**, p. 861, and 1954, **93**, p. 1204.
(11) "Transistor Teachers' Summer School," *ibid.*, 1952, **88**, p. 1368.
(12) SHOCKLEY, W.: *Bell System Technical Journal*, 1949, **28**, p. 435.
(13) GOUCHER, F. S., PEARSON, G. L., SPARKS, M., TEAL, G. K., and SHOCKLEY, W.: *Physical Review*, 1951, **81**, p. 637.
(14) PFANN, W. G.: *Transactions of the American Institute of Mining Engineers*, 1952, **194**, p. 747.
(15) PFANN, W. G., and OLSEN, K. M.: *Physical Review*, 1953, **89**, p. 322.
(16) PFANN, W. G., and OLSEN, K. M.: *Bell Laboratories Record*, 1955, **33**, p. 201.
(17) TEAL, G. K., and LITTLE, J. B.: *Physical Review*, 1950, **78**, p. 647.
(18) BURGERS, J. M.: *Proceedings of the Physical Society*, 1940, **52**, p. 23.
(19) BURGERS, J. M.: *Proceedings Koninkijke Nederlandsche Akademie van Wetenschappen*, 1939, **42**, p. 293.
(20) BRAGG, W. L.: *Proceedings of the Physical Society*, 1940, **52**, p. 54.
(21) VOGEL, F. L., PFANN, W. G., COREY, H. E., and THOMAS, E. E.: *Physical Review*, 1953, **90**, p. 489.
(22) SHOCKLEY, W., and READ, W. T.: *ibid.*, 1949, **75**, p. 692.
(23) TEAL, G. K., SPARKS, M., and BUEHLER, E.: *ibid.*, 1951, **81**, p. 637.
(24) HALL, R. N., and DUNLAP, W. C.: *ibid.*, 1950, **80**, p. 467.
(25) HALL, R. N.: *Proceedings of the Institute of Radio Engineers*, 1952, **40**, p. 1512.
(26) SABY, J. S.: *ibid.*, p. 1358.
(27) FULLER, C. S., and DITZENBERGER, J. A.: *Journal of Applied Physics*, 1954, **25**, p. 1439.
(28) PEARSON, G. L., and FULLER, C. S.: *Proceedings of the Institute of Radio Engineers*, 1954, **42**, p. 760.
(29) CHAPIN, D. M., FULLER, C. S., and PEARSON, G. L.: *Journal of Applied Physics*, 1954, **25**, p. 676.
(30) CHAPIN, D. M., FULLER, C. S., and PEARSON, G. L.: *Bell Laboratories Record*, 1955, **33**, p. 241.
(31) SHOCKLEY, W., SPARKS, M., and TEAL, G. K.: *Physical Review*, 1951, **83**, p. 151.
(32) PIETENPOL, W. J., and WALLACE, R. L.: *Proceedings of the Institute of Radio Engineers*, 1951, **39**, p. 753.
(33) GEBALLE, T. H., and MORIN, F. J.: *Physical Review*, 1954, **95**, p. 1085.
(34) MORIN, F. J., MAITA, J. P., SHULMAN, R. G., and HANNAY, N. B.: *ibid.*, 1954, **96**, p. 833.
(35) BURTON, J. A.: *Physica*, 1954, **20**, p. 845.
(36) HORNBECK, J. A., and HAYNES, J. R.: *Physical Review*, 1955, **97**, p. 311.
(37) SHOCKLEY, W.: *ibid.*, 1953, **91**, p. 215.
(38) DRESSELHAUS, G., KIP, A. F., and KITTEL, C.: *ibid.*, 1953, **92**, p. 827.
(39) DEXTER, R. N., LAX, B., KIP, A. F., and DRESSELHAUS, G.: *ibid.*, 1954, **96**, p. 222.
(40) DEXTER, R. N., and LAX, B.: *ibid.*, p. 223.
(41) LAX, B., and MAVROIDES, J. G.: *ibid.* (to be published).
(42) DEXTER, R. N., ZEIGER, H. J., and LAX, B.: *ibid.* (to be published).

621.398 : 681.142

Paper No. 1905 M
Jan. 1956

A TRANSDUCER FOR DIGITAL DATA-TRANSMISSION SYSTEMS

By R. H. BARKER, B.Sc., Ph.D., Associate Member.

(*The paper was first received* 21*st December*, 1954, *and in revised form* 3*rd June*, 1955.)

SUMMARY

Accurate information such as may be required for weapon control may best be transmitted over long-distance telegraph or voice communication circuits by using some form of digital representation or pulse code modulation. The paper describes a method by which the position of a scale attached to a datum shaft may be "read" photoelectrically as a binary number in much the same way as the scale of a surveying instrument is read in degrees and minutes. Data-transducers have been constructed to represent the angular position of a shaft as a 14-digit binary number, that is to an accuracy of rather better than one minute of arc. Special precautions have been taken in the design of the scale to prevent gross errors due to movement of the scale while it is being read. Such errors as do occur are examined in detail, since in some data-transmission applications they may not be sufficiently serious to justify the more complex equipment required for their complete elimination.

(1) INTRODUCTION

In its widest sense the term "data transmission" implies the transmission from one place to another of some kind of factual information. Nevertheless the expression has a peculiar and restricted field of use derived from its original use in connection with the control of guns. The development of visual, acoustic and radar aids for the determination of target positions has led to the need for suitable equipment for the transmission of the data so obtained to the guns or their predictors. In the terminology of the fighting Services the expression embraces the transmission of target co-ordinate data for any form of weapon control.

Apart from its original field of application there are certain other considerations which serve to define data transmission as a particular branch of the much wider subject of information transmission. The most important one is that of quantitative accuracy. The accuracy with which the data are transmitted must be capable of precise specification, and it is desirable that the errors of transmission should be significantly less than the errors of measurement.

In attempting to define the boundaries of the subject it is necessary to take note of and to exclude the closely allied subject of telemetry. In the latter, as the word implies, measurements have to be made at a distance, whereas in the former the measurements have presumably already been made and only the results are to be transmitted. This difference is more apparent than real. For example, the bearing of an aircraft may be measured by radar, the result being available locally as the angular position of a shaft. When the result is transmitted to a distant place, should the process be called telemetry of shaft position or angular data transmission?

A study of the historical associations reveals a distinction not implied in the nomenclature. The survey made by Whitehead and Walsh,[1] for example, shows that telemetered data are almost invariably recorded for future analysis. In what are generally classed as data transmission systems, on the other hand, the received information is usually, though not invariably, reconstituted into its original form. Moreover it is obvious that where weapon control is the objective, this must be accomplished with the least possible delay.

In fire-control applications the data are analogues of target co-ordinates or some function thereof. A very common analogue is the angular position of a shaft and this was used in most of the fire-control radars and predictors built during the Second World War. In a review by Bell[2] of systems used by the fighting Services, the input data are in all cases presented as shaft rotation, and it is probable that this particular analogue will remain important for some time to come.

Except in rare instances the distances over which data were transmitted during the War were short enough for the transmitters and receivers to be connected together by the necessary number of wires. This can hardly be said to pose a problem in communications. With the advent of high-speed aircraft and guided missiles, however, larger distances are likely to be involved and the communication link may be of paramount importance. New types of data transducers will be required to provide electrical signals whose accuracy will not suffer on the way. The form of modulation selected should be such as to minimize the effects of added noise and of amplitude, frequency and phase distortion.

In this respect the digital representation of the data has many advantages, not the least of which is that the signal may be regenerated with an extremely low probability of error, provided that the noise and distortion are kept within certain reasonable limits. Furthermore this process may be repeated as many times as are necessary to cover the distance involved. For such a digital representation all possible and acceptable values of the data on a numerical scale are represented by separate patterns of pulses, or codes. Hence the term "pulse code modulation" is usually applied.

Each code group represents to the nearest unit the value of the input quantity at a particular instant and constitutes a sample. A well-known result of communication theory is that a signal having an upper frequency f_c may be accurately reconstructed from samples taken at a frequency not less than $2f_c$. It is often forgotten, however, that this involves a delay, perhaps of many sampling intervals, and it is not a satisfactory rule of thumb to apply to a data transmission system.

It is important with long-distance communication circuits to economize as much as possible in bandwidth and hence, for this application, in the number of samples sent per second. The significant consequence with regard to the design of the data transducer is that the input quantity may change considerably during the time taken to transmit the code group representing one sample. If the coding were performed instantaneously the group would refer accurately to that instant, but storage facilities would have to be provided to permit the required slow subsequent transmission in a very restricted frequency band. On the other hand such storage may be avoided by devising a transducer in which the coding is a continuous process proceeding synchronously with the transmission, but only at the expense of errors when the input quantity is changing rapidly.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
Dr Barker is at the Ministry of Supply.

he main purpose of the paper is to describe a transducer (of latter type) in which such errors are minimized, and to form ssessment of the magnitude of these errors. In a companion er[3] it is shown that they are of the same order of magnitude hose inherent in reconstituting the data with no delay at the iver. Fig. 1 shows a typical digital data-transmission system

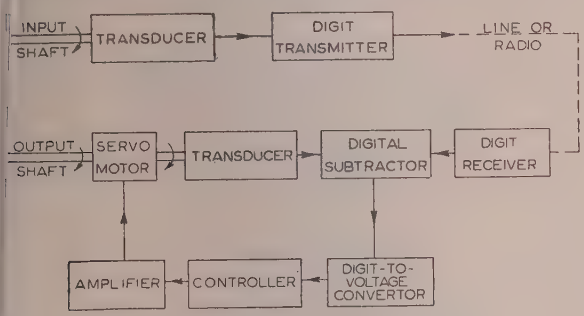


Fig. 1.—A digital data-transmission system.

d illustrates the part played by the transducer. The initial put, in the form of a shaft position, is converted into pulse de modulation of, say, 10 or more binary digits per group. ansmission is at the rate of 50 bits/sec over a normal teleaph circuit. At the receiver the demands for accuracy in constitution may be more readily met by making use of a rvo system. The output shaft carries a second similar transducer e output of which is subtracted (digitally) from the incoming gnal. The difference is the error and is converted into a rresponding voltage for use in controlling the servo motor.

(2) INDIRECT METHODS OF CODING

A variety of methods is now available for the digital encoding a variable voltage. Oxford,[4] for example, in 1947 was one the first to devise a simple system using only a few valves of nventional type. Since then numerous workers have made provements, and it is now possible to encode to an accuracy about 0·2 volt a voltage varying over a range of ± 100 volts: at is to say, to an accuracy of one part in a thousand or to n binary digits. Potentiometers of similar accuracy are ailable commercially so that encoding of shaft position to an curacy of 0·1% should be possible. It is not at present ossible, however, to permit continuous rotation without a gap, the applications for such a system are limited.

An alternative approach is to encode the electrical output of Magslip transmitter. Allard and Hill[5] have described a special pe of cathode-ray tube which has been developed for this rpose in which a transverse magnetic field is caused to rotate ith the datum shaft, just as it does in a Magslip receiver. stead of the conventional screen, the tube is fitted with an sembly of sector-shaped plates from which digital reprentation of the angle may be obtained.

(3) USE OF CALIBRATED SCALE

The use of a calibrated scale with a suitable reading device r the measurement of angle is by no means new. In the best rveying instruments, for example, an accurately graduated ale is read through a magnifying eyepiece to an accuracy of a w seconds of arc. Whether or not to incorporate a gear train improve the accuracy or resolution of an inherently less curate device is a matter for consideration in individual cases. the requirement can be met, however, with a single scale on e datum shaft it is probably better to do so. The scales can be reproduced photographically from a master and do not suffer from wear. The problem is to devise a scale which can be read electrically as a train of digital pulses instead of with the human eye.

It is a problem which came to the notice, quite independently, of both the author and W. S. Elliott[6] during the latter half of 1948, and both, still independently, came to the conclusion that the only practicable solution was to produce the scale photographically on a glass disc and to "read" the position by means of a cathode-ray-tube light source and a photocell.

Fig. 2 is a diagram of a circular scale constructed according

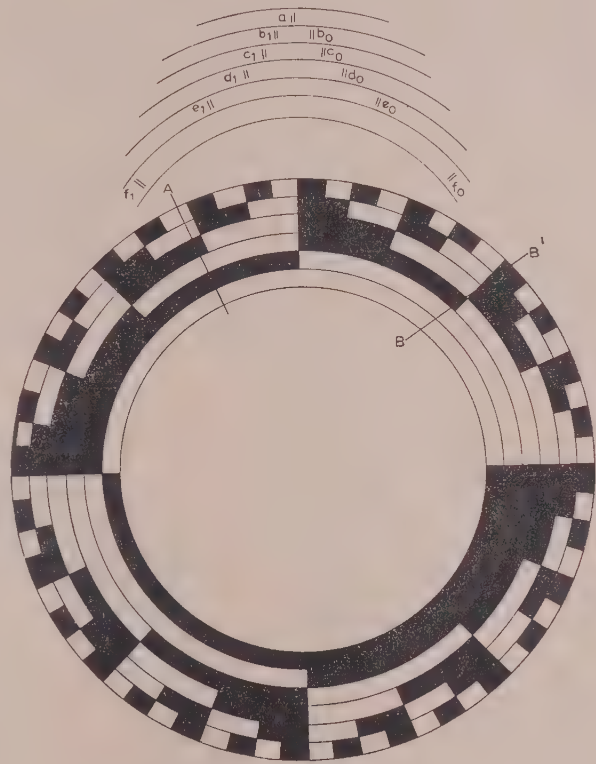


Fig. 2.—Circular binary pattern.
6-digit, illustrating a source of error.

to the binary code. The complete binary number denoting a particular angle is obtained by scanning the pattern radially, as at the index mark A. The white sections will be arbitrarily taken as 1's and the black sections as 0's. Thus the angle at A is denoted by the binary number .101011 in which the most significant digit (that following the binary point) is read from the inside ring and has the significance $\frac{1}{2}$. Successive digits have significance $\frac{1}{4}$, $\frac{1}{8}$, etc., so that position A corresponds in decimal notation to 43/64. Fig. 2 also shows in an exaggerated manner an important potential source of error and one which must be eliminated at all costs. Owing to inaccurate scale manufacture, to inaccurate radial adjustment of the fiduciary slit or to movement of the scale during reading, it is possible that the effective scanning line may be such as BB′. The indicated number would be .110000, which corresponds actually to an angle 45° anticlockwise from B. Since in the scales ultimately to be used the smallest element is approximately 0·001 in wide it is obvious that special measures are necessary for avoiding this type of error.

(3.1) The Elliott Method

If the amount of tolerance to be provided can be made sufficiently small the method described by Elliott, Robbins and Evans[6] is very satisfactory and straightforward. The order of scanning is with the least significant digit first. Ahead of the first digit at every position where this particular type of error can occur is a white window. A scanning spot which does not pass one of these windows is caused to be deflected a small amount to one side where it can traverse the pattern clear of the uncertainty associated with the vertical edges. The margin provided by this method is of course small and very accurate alignment of scanning direction is still required. Moreover, the scale must not move more than a small fraction of the finest digit during the time taken to scan across. The scanning rate is very high in Elliott's application, being linked to a high-speed digital computer, and so does not constitute a significant limitation.

(3.2) The Cyclic Permuting Code

An alternative method has been in use for some time and is now fairly well known. It involves a variation of the binary scale known as the cyclic permuting (C.P.) code—or in America as the Gray code. Any binary number may be converted digit by digit to the C.P. code by the following rule: those binary digits for which the next more significant digit is zero remain unchanged; the others are altered, 1 into 0 and 0 into 1. Examples are shown in Table 1.

Table 1

THE C.P. CODE

Decimal notation	Binary scale	C.P. code
0	0000	0000
1	0001	0001
2	0010	0011
3	0011	0010
4	0100	0110
5	0101	0111
6	0110	0101
7	0111	0100
8	1000	1100
9	1001	1101
10	1010	1111

The important property of the C.P. code is that only one digit changes at a time. The pattern obtained by using this code is as shown in Fig. 3(*a*).

As an aside it is interesting to note that the cyclic permuting variant is not confined to binary numbers. The more general rule is that any digit remains unchanged if the number formed by those digits of greater significance is even, otherwise it is replaced by its complement. In this context the complement of a digit is obtained by subtracting the actual value of the digit from its maximum possible value. Thus the C.P. version of 2142 on the quinary scale is 2102. The second digit, 1, remains unchanged because 2 is even. The third digit, 4, is replaced by its complement, 0, because 21 (in decimal notation 11) is odd. The fourth digit, 2, is replaced by its complement, also 2, because 214 (in decimal notation 59) is odd.

Refer again to the binary cyclic permuting pattern of Fig. 3(*a*), in which a number of possible lines of scan are indicated, corresponding to various rates of movement of the scale during scan. These are all drawn starting from a point just to the left of where the most significant digit changes and are labelled a, b, c, etc. The most significant digit, according to the convention

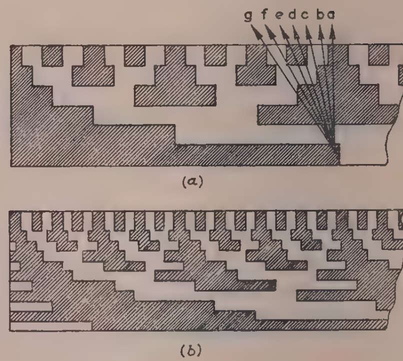


Fig. 3.—Modified binary patterns.
(*a*) Cyclic permuting code.
(*b*) V-scan pattern.

adopted, is recorded as zero. The lines of scan, however, migh equally well have been drawn from a point just to the right o this, when the most significant digit would have been unity Although they are not drawn in, such lines are referred to a a′, b′, c′, etc., in Table 2, which shows the numbers as actuall read.

If the true position is defined as that corresponding to th

Table 2

EXAMPLES OF SCANNING ERRORS

Line of scan	Movement in 64ths	C.P. code	Binary scale	Number in 64ths
a	$\frac{1}{2}$	.010000	.011111	31
a′		.110000	.100000	32
b	$1\frac{1}{2}$	.010001	.011110	30
b′		.110001	.100001	33
c	$2\frac{1}{2}$	.010001	.011110	30
c′		.110001	.100001	33
d	$3\frac{1}{2}$	.010010	.011100	28
d′		.110010	.100011	35
e	$4\frac{1}{2}$	.010010	.011100	28
e′		.110010	.100011	35
f	$5\frac{1}{2}$	.010011	.011101	29
f′		.110011	.100010	34
g	$6\frac{1}{2}$	.010111	.011010	26
g′		.110111	.100101	37

instant at which the most significant digit is read, the maximur error in these examples is 5/64, but if it is taken as the instan at which the least significant digit is read the error may be muc larger. The indicated number may in fact lie outside the rang of true position through which the scale moves during the perio of scan. This example has been chosen to illustrate that quit large errors may be obtained. At other points on the scale th errors will often be less. A more general study shows that, i the movement during scan is θ, and the number of digits is large the possible error may approach $\pm\theta$ or 2θ depending upo whether the instant of reading the most or the least significan digit is taken as the instant of reference.

The conversion of a C.P. code number back to the "straight" form must proceed starting with the most significant digit, whic is unchanged. Taking the remaining digits in turn, any one o them remains unchanged if the number (straight) formed by th

gits of greater significance is even; otherwise it is replaced by ; complement. This results in a particularly simple operation ith binary numbers. Starting with the most significant digit e number of 1's is counted up to and including the digit in iestion. If the result is odd the digit becomes 1, otherwise it comes 0.

A further feature which requires discussion in relation to the .P. code is the operation of subtraction which must be sub-quently carried out in deriving the error signal. A system of ithmetic has not yet been evolved which is applicable to umbers in the C.P. code, so that conversion to the straight form necessary. This must proceed with the most significant digit st. The digital subtraction, on the other hand, must proceed ith the least significant digit first since the carry figure is always ansferred to a place of greater significance. The complication of ving to arrange for a reversal of the digit sequence would be nsiderable. Professor D. R. Hartree has pointed out in private rrespondence that the complete conversion to the straight nary scale may be avoided if the parity (oddness or evenness) each of the numbers be known. The parity of the number is the me as that of the number of 1's in the C.P. code form. It may indicated by prefixing an additional digit, 1 if the parity is odd, f it is even. This further digit might perhaps be incorporated in e scale itself. Let the parity digit be represented by A_0 and the P. code digits by $A_1, A_2 \ldots A_n$, A_1 being the least significant git and coming first in time after A_0. Conversion digit by git is now possible, starting with A_1. Let $a_1, a_2 \ldots a_n$ be the corresponding digits on the binary scale. Hartree has shown that the digit a_{k+1} is given by the parity, with the above representation, of the A_k and a_k digits. Thus a_{k+1} is defined by Table 3.

Table 3

A_k	0	0	1	1
a_k	0	1	0	1
a_{k+1}	0	1	1	0

The operation of determining the parity of two digits in this manner is also known as modulus 2 addition. This method of dealing with C.P. coded numbers has been mentioned because it is not generally known and may have some applications. So far as the present work is concerned, no way has been found of including the parity digit upon the scale, and the alternative of counting the number of 1's in the C.P. code form involves a delay which is undesirable and would also necessitate storage of the number during the count.

The advantages of the C.P.-code method of avoiding scanning errors will be summarized. It results in a simple pattern plate requiring no special control of the scanning spot. The widths of the black and white sections are doubled compared to the straight scale, so making for better resolution. Adequate tolerances are provided for scale or scanning inaccuracies or scale movement, and gross errors are completely avoided. Even larger tolerances are, however, provided in the method to be described later. The only significant disadvantage for the present application is that the conversion to the straight binary scale does not integrate well with the arithmetic subtraction.

Fig. 4.—10-digit V-scan pattern.

(3.3) The V-Scan Method

As explained above, it is highly desirable that the order of digits generated by the transducer shall be with that of least significance first, and the principal object in developing the method now to be described[7] was to combine this property with greatest possible tolerance as regards scale movement. The basis of the method is that each digit after the first may be read in one of two alternative ways, chosen according to the value of the preceding (less significant) digit. Refer again to Fig. 2 and imagine the arrangement of eleven slits shown at the top to be actually superimposed upon the circles of the pattern so that slit (a) lies on top of the outside ring. These slits serve to define possible positions for the scanning spot. Slit (a) is first illuminated. If the corresponding digit is a 1 (white according to the assumed convention) then, for the next digit, slit b_1 on the left

is illuminated, but if the first digit should be 0 slit b_0 would next be illuminated instead. Similarly, for the third digit either slit c_1 or slit c_0 is illuminated according to whether the second digit was 1 or 0, and so on. The displacement accorded to a pair of slits should be equal to one element of the ring previously scanned, though normally much less than this will suffice for digits of high significance. The available tolerance for scanning errors is shown in Table 4 in which n is the number of digits.

If the rate of scan is x digits/sec and other sources of error are neglected, an angular velocity of the scale of $2^{-n}\pi x$ rad/s will just cause reading errors. Two experimental systems have been built. In the first, having 10 digits and a scanning rate of 50 bits/sec, this limiting velocity is 8 deg/sec. In the second, having 14 digits and a scanning rate of 1 500 bits/sec, it is nearly 17 deg/sec. Incidentally, it is the signal from the former that can be transmitted over a telegraph circuit; the latter requires a speech circuit.

The system of slits in Fig. 2 is a little inconvenient to apply in practice and a variation has been adopted. Instead of providing two alternative angular positions at which each ring of the scale may be examined, the scan takes place along a radius and each ring (after the first) is split into two half-rings having the displacements quoted previously. The appropriate half-ring is selected according to the value of the preceding less significant digit. The form of the pattern is as shown in Fig. 3(*b*). Fig. 4 is a reproduction of the 10-digit scale used in the subsequent experimental work.

A further important feature of the V-scan type of pattern plate is that even with rates of rotation such that the tolerances

Table 4

Tolerances Provided by the V-Scan Method

Order of ring	Spacing of slits	Total tolerance
$n-1$	$\pi/2^{n-1}$	$\pi/2^n$
$n-2$	$\pi/2^{n-2}$	$\pi/2^{n-1}$
.	.	.
.	.	.
.	.	.
2	$\pi/4$	$\pi/8$
1	$\pi/2$	$\pi/4$

listed in Table 4 are excèeded the reading errors are still not excessive. An attempt has been made to estimate their magnitudes in certain special cases, but rather drastic simplifying assumptions have to be made in order to reduce the problem to one of reasonable magnitude. At the start of the scan the spot is taken to be positioned either exactly over one element or exactly half-way between two adjacent ones, and both these cases are equally likely. This assumption is equivalent to quantizing the angular position of the scale in half-element steps. Since the width of the scanning spot is likely to be of the same order as that of the element, this is not unreasonable.

(3.3.1) Example 1: Scale Stationary.

With the true scale position as an exact integer, the spot lies exactly behind an element and there is no reading error. With the true scale position an odd half-integer, the spot lies on a black/white boundary and the generated digit is equally likely to be 1 or 0. The reading error is therefore $-\frac{1}{2}$ or $+\frac{1}{2}$. The nature of the pattern is such that subsequent digits will not be in error and the overall assessment may be expressed as a table of probabilities, thus:

Table 5

Amount of error . .	$-\frac{1}{2}$	0	$+\frac{1}{2}$
Probability . .	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

The average error is zero.

(3.3.2) Example 2: Rotation at $\frac{1}{2}$ Element per Digit.

It is convenient to express the rate of rotation as the number of elements through which the scale moves in the time allotted to one digit. A speed of $\frac{1}{2}$ element per digit corresponds to 8 deg/s in the 10-digit equipment. At this speed there will still be no reading error when the true position (referred to the time at which the least significant digit is read) is an exact integer. If it is an odd half-integer the spot may be positioned on the black/white boundary of the second and even perhaps the third digit. The method of assessing the errors is indicated in Table 6 in which all the alternatives are shown. The binary numbers are written as usual with the least significant

Table 6

Calculation of Error Probabilities

Elements per digit	True position	Number generated	Error (lead)	Summary: Error	Summary: Probability	Summary: Average
0	0	. . . 0 0 0 0 0	0	$-\frac{1}{2}$	$\frac{1}{4}$	
	$\frac{1}{2}$	. . . 0 0 0 0 0	$-\frac{1}{2}$	0	$\frac{1}{2}$	0
		. . . 0 0 0 0 1	$+\frac{1}{2}$	$+\frac{1}{2}$	$\frac{1}{4}$	
$\frac{1}{2}$	0	. . . 0 0 0 0 0	0	$-\frac{1}{2}$	3/32	
	$\frac{1}{2}$	. . . 0 0 0 0 >0	$-\frac{1}{2}$	0	$\frac{1}{2}$	
		. . . 0 0 0 1	$+1\frac{1}{2}$			
		. . . 0 0 0 0 1	$+\frac{1}{2}$	$+\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{8}$
	1	. . . 0 0 0 0 1	0	$+1\frac{1}{2}$	$\frac{1}{8}$	
	$1\frac{1}{2}$	. . . 0 0 0 1 0	$+\frac{1}{2}$	$+3\frac{1}{2}$	1/32	
		. . . 0 0 0 >0	$-\frac{1}{2}$			
		. . . 0 0 1 >1	$+3\frac{1}{2}$			
		. . . 0 0 0 1				
1	0	. . . 0 0 0 0 0	0			
		. . . 0 0 0 1 0	2	0	$\frac{1}{8}$	
	$\frac{1}{2}$	. . . 0 0 0 1 0	$1\frac{1}{2}$	$\frac{1}{2}$	3/16	
		. . . 0 0 0 >0 1	$\frac{1}{2}$			
		. . . 0 0 1	$4\frac{1}{2}$	$1\frac{1}{2}$	$\frac{1}{4}$	
	1	. . . 0 0 1 0 1	4	2	$\frac{1}{4}$	$1\frac{7}{8}$
		. . . 0 0 0 1 1	2			
	$1\frac{1}{2}$	. . . 0 0 0 1 0	$\frac{1}{2}$	4	$\frac{1}{8}$	
		. . . 0 0 0 1 1	$1\frac{1}{2}$	$4\frac{1}{2}$	1/32	
	2	. . . 0 0 1 0 0	2	$8\frac{1}{2}$	1/32	
		. . . 0 0 0 1 0	0			
	$2\frac{1}{2}$	. . . 0 0 1 0 0	$1\frac{1}{2}$			
		. . . 0 0 >0 1 1	$\frac{1}{2}$			
		. . . 0 1	$8\frac{1}{2}$			
	3	. . . 0 0 1 0 1	2			
		. . . 0 0 1 1 1	4			
	$3\frac{1}{2}$	. . . 0 0 1 0 0	$\frac{1}{2}$			
		. . . 0 0 1 0 1	$1\frac{1}{2}$			

it on the right, and the order in which the digits are generated herefore from right to left. The entry $\overset{0}{\underset{1}{}}>0$ means that the st significant digit is 0 and that the next digit is in consequence ually likely to be 0 or 1. Only four entries are necessary at speed ½ element per digit since the sequence of errors repeats reafter. The error probabilities may therefore be evaluated te simply.

Table 6 also includes the case in which the speed is 1 element digit. The calculations have in fact been extended up to lements per digit, but are too involved to be reproduced here.

Table 7

Summary of Calculated Reading Errors

Speed			Errors (leading)			
lements er digit	Degrees per second	Elements per scan	Minimum	Maximum	Average	R.M.S.
0	0	0	$-\frac{1}{2}$	$\frac{1}{2}$	0	0·35
½	8	5½	$-\frac{1}{2}$	3½	0·37	0·78
1	16	11	0	8½	1·87	1·73
1½	24	16½	½	10	3·25	2·51
2	32	22	1½	18½	5·87	3·00
2½	40	27½	2½	21½	6·48	3·59
3	48	33	3½	23½	9·75	5·08

tead the results are summarized in Table 7. It will be noted t the average error is always in the leading direction and that maximum error is always less than the amount by which the le moves during the time of scan. At speeds of one or more ments per digit the actual error is always in the leading ection. It must be emphasized that this assessment of errors necessarily very crude, being based upon a drastic overplification of the actual conditions. Nevertheless, it does icate that if excessive speed should develop, as during slewing, numbers generated will still be reasonably correct.

(4) EXPERIMENTAL 10-DIGIT SCALE READER

The data-conversion equipment described in this Section was nstructed for use in an experimental digital servo-system of type illustrated in Fig. 1. The number of digits (ten) was cted as being the largest for which it was practicable to w the scale by hand. The rate of scanning, 50 bits/sec, was cted to be suited to transmission over telegraph circuits. In case, a high rate of scan would tend to reduce some of the ors which it is desirable to investigate.

The scale of Fig. 4 was originally drawn approximately 2 ft diameter. It was reduced photographically to just under n in diameter on a glass disc mounted upon the datum ft. Immediately behind the disc and almost touching it the cathode-ray tube used for scanning, and about 6 in ay on the opposite side is a multiplier type of photoelectric . This gives an output whenever the scanning spot is not cured by opaque parts of the scale.

n a digital equipment in which the various digits occur ally, the only method of identifying a particular digit so t its significance is known is by the instant at which it occurs. s entails the synchronous operation of the various parts of system, namely transmitting and receiving data-convertors, tractor, error-convertor, etc. Synchronism is ensured by provision of a special timing circuit, the essential features of ch may be appreciated by reference to Fig. 5, which shows ne idealized waveforms. At Fig. 5(*a*) is a 50-c/s square wave, negative-going edges of which define the separation of one t from the next. Successive digits are counted by means of

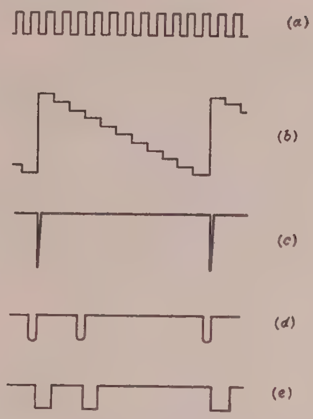


Fig. 5.—Idealized scanning waveforms.
(*a*) Basic timing waveform.
(*b*) Staircase scanning waveform.
(*c*) Synchronizing pulses.
(*d*) Photocell output.
(*e*) Delayed output.

an energy-storage type of counter in which the waveform at Fig. 5(*b*) appears. The top step of this waveform identifies the least significant digit. There are eleven steps, of which ten define the ten digits of the number, and the eleventh, referred to as the synchronizing or S-digit, serves a variety of purposes to which attention will be drawn later. The flyback of the staircase waveform is accompanied by a large negative pulse, shown at Fig. 5(*c*), which is used to reset the counters in other units of the equipment. To avoid confusion it is well to explain that the top step of the waveform at Fig. 5(*b*) identifies the least significant digit only in the circuit of the photocell and its amplifier. Subsequent operations involve delay, and in later stages the least significant digit may occur at the time of the second or even the third step. Movement of the spot radially is by a series of jumps, the spot dwelling for the period of each digit behind the corresponding ring of the scale. The waveform inducing this movement is as shown in Fig. 5(*b*) and results in a cathode-ray-tube display consisting of a row of eleven dots (actually consecutive in time) which, by means of amplitude and shift controls, are positioned accurately behind the outer halves of each ring of the pattern. In addition, the spots are blacked out for the first half of each digit period in order to avoid the need for excessively large time-constants in the amplifiers.

The waveform in Fig. 5(*d*) is typical of the signal obtained from the photocell. Pulses are always obtained for the S-digit and a single pulse is shown in position 3, thereby indicating the number 4. The function of the control circuit is to cause the scanning spot to jump to an inner ring instead of an outer ring if the preceding digit was a 1, i.e. if an output pulse was obtained from the photocell. The first stage is an amplitude discriminator, or slicer, which gives a definite output if, and only if, the pulse exceeds a certain level. The slicer output is regenerated as a full-width pulse in the following digit period, the waveform being shown in Fig. 5(*e*). A corresponding current passes through the deflection coils of the cathode-ray tube and is responsible for the incremental deflection which causes the spot to explore the inner half-ring instead of the outer half-ring. The waveform in Fig. 5(*e*) also constitutes the output signal of the transducer. Note that it is delayed with respect to the photocell waveform so that the beginning of the S-digit is now timed to coincide with the synchronizing pulse of Fig. 5(*c*). A refinement that has been found desirable consists of an automatic brilliance-control circuit of the cathode-ray tube. In effect the peak voltage of the photocell signal is measured, smoothed and

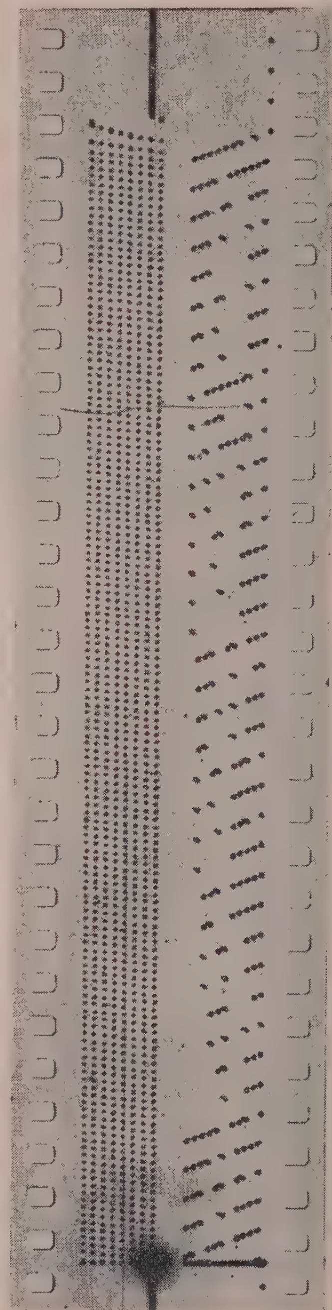
applied as negative feedback to the beam-control electrode of the tube. One of the reasons for the provision of the S-digit is that, without it, there would be no indication of brilliance when the disc is set in position 0. In addition, without it, it would be necessary to use d.c. coupling between circuits in order to distinguish between positions 0 and 1023.

The most important function of the S-digit, however, is in the provision of a synchronizing signal over a long-distance communication circuit. The method consists of alternating the S-digit between 1 and 0 in each successive frame. At the receiver, a test circuit explores successive digits until it finds one which continues to alternate regularly in this manner, so establishing correct synchronization.

(4.1) Dynamic Reading Errors

The experimental determination of reading errors necessitates an independent method of knowing the true position of the scale. A simple expedient is to arrange a lamp and photocell opposite the least-significant-digit ring and to use it to count successive elements as the scale rotates. The photocell output is amplified and limited and used to drive a two-stage binary counter. By combining in suitable proportions the outputs of the amplifier and the two binary stages it is possible to obtain a signal which has eight permitted levels. Such a signal was recorded on moving film alongside the actual numbers read from the scale. The accompanying oscillogram, Fig. 6, is typical of the results and was taken at a rotation speed of 49 deg/sec. On the left-hand side is the output of the modulus 8 counter. The initial level is random. On the right-hand side are the binary numbers displayed according to the convention that a dot present indicates 1 and a dot absent indicates 0. The S-digit is on the extreme right and the least significant next to it. Owing to the one-digit delay in the scale-reading equipment the S-digit was recorded at the instant that the least significant digit was read. Thus the true scale position is determined by the coincidence in time (i.e. in vertical level on the oscillogram) of the S-digit and one level of the modulus 8 counter. A reference is provided by always starting the scale from rest in the zero position.

Reading from the film record becomes easy with practice, though laborious. It is facilitated by working on the scale of 8 instead of 10. The ten binary digits are divided into groups of 1, 3, 3, 3. Each group of three may be written as one of the digits 0 to 7. This has been done in the centre column of figures on the oscillogram. The output of the modulus 8 counter is similarly expressed in the left-hand column. The true-position number increases as the scale-reading number decreases. If there were no error the total would remain constant. The actual error (leading) is the amount by which the total exceeds the initial setting of the modulus 8 counter. Note that the arithmetic and the tabulation of errors are also carried out in the scale of 8. The results may be plotted graphically without



True position	Scale reading	Error (lead)
6	0	0
12	1772	2
41	1737	6
76	1663	25
135	1642	7
174	1603	7
234	1542	10
274	1503	7
332	1443	11
372	1376	16
432	1342	12
472	1276	16
530	1245	11
571	1206	7
627	1146	11
670	1107	7
727	1047	10
766	1007	11
1025	733	26
1065	711	10
1124	653	7
1164	613	7
1224	556	6
1263	517	6
1322	456	6
1362	417	7
1420	351	7
1461	317	6
1517	263	6
1556	223	5
1616	161	7
1655	122	7
1715	63	6
1754	23	7
13	1756	15
53	1727	6
111	1666	7
150	1627	7
207	1567	10

Fig. 6.—Oscillographic record of dynamic reading errors.

Table 8

Experimentally Determined Errors

Speed		No. of readings	Errors (leading)		
Degrees per second	Elements per scan		Minimum	Maximum	Average
3·4	2·34	484	$-1\frac{1}{2}$	$+1\frac{1}{2}$	$-0·1$
9·45	6·5	174	-1	$+3$	0·15
14·6	10	112	-1	5	1·45
23·0	15·8	71	$+1$	11	3·4
33·4	23	49	3	18	6·8
49	33·6	36	5	22	8·5

conversion using paper ruled 8 lines to the inch. Fig. 7 shows results so obtained for six different rates of rotation.

An attempt was made, with the record taken at 3·4 deg/sec to estimate the error to the nearest half-element. Since the speed is well below $\frac{1}{2}$ element per digit, one would not expect to find errors exceeding half an element. In fact the errors extend to $\pm 1\frac{1}{2}$ elements. The general shape of the graph shows, however, that there is a systematic error, of the order of $\pm\frac{1}{2}$ element in addition to the normal reading error of $\pm\frac{1}{2}$ element. The systematic error was found to be due partly to dis-

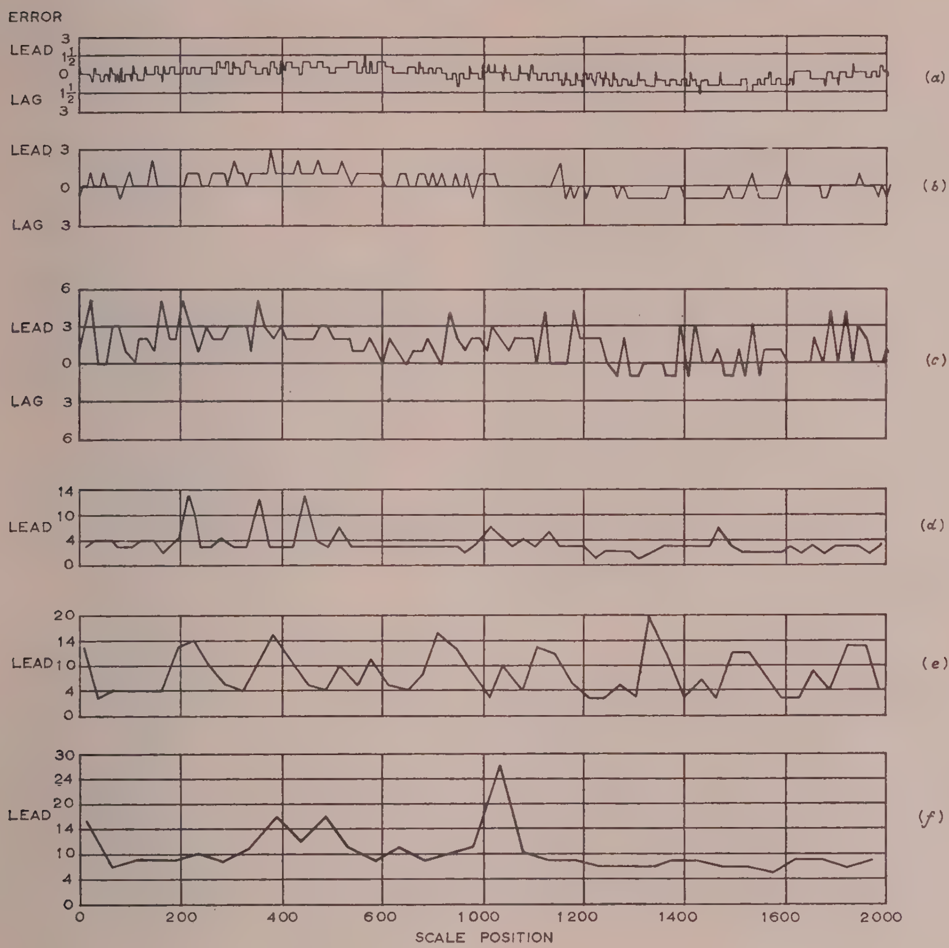


Fig. 7.—Dynamic reading errors.

(a) Speed 3·4 deg/s.
(b) Speed 9·4 deg/s.
(c) Speed 14·6 deg/s.
(d) Speed 23 deg/s.
(e) Speed 33·4 deg/s.
(f) Speed 49 deg/s.

ortion of the intermediate film used in the manufacture of the attern plate, and partly to inaccurate centring. It is apparent ecause the lamp and photocell assembly used for the true-osition determination is mounted diametrically opposite to the canning unit. The systematic error can also be detected in the ther two curves (9·4 and 14·6 deg/sec) of Fig. 6. The experi-nental results are summarized in Table 8. Ordinary decimal otation is used.

Fig. 8 enables a direct comparison of theoretical and experi-nental results to be made. The agreement is very close, specially when allowance is made for the systematic errors hich noticeably affect the results at the slower rates of rotation. further source of systematic error is that the initial position uncertain to the extent of $\pm\frac{1}{2}$ element.

(5) HIGH-ACCURACY (14-DIGIT) SCALE READER

The principal *raison d'être* of the pattern-plate type of data-onvertor is that the errors and wear associated with gears of easonable dimensions are eliminated, and there is little point using one unless very high accuracy is required. An attempt as therefore made to produce photographically scales of the type described in Section 4 but reading to an accuracy of about 1 minute of arc. Scales were constructed with 14 binary digits on which the maximum error of positioning is only a few seconds of arc. With a servo system in which scales copied from the same master are used both at the transmitter and receiver, systematic errors will tend to cancel out. The inherent accuracy of the scales is so high, however, that no significant advantage could thereby be gained.

(5.1) Production of Accurate Scales

A technique for making high-accuracy scales was developed during 1950 by a commercial firm having previous experience of manufacturing scales for optical instruments by photographic means instead of by engraving. In their method, a photographic plate is placed upon the table of an automatic circular dividing engine, and a small section of scale is projected upon the plate for the duration of the required exposure. The dividing engine then turns the plate through an angle exactly equal to the arc of scale first projected, and the process is repeated. The section of scale normally consists of one or two degrees divided into minutes or half-minutes. Numbering of the degrees is done by means of a second projector in which a film strip carries numbers

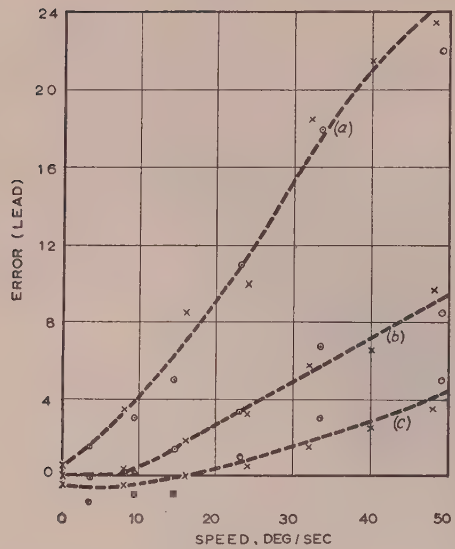


Fig. 8.—Comparison of errors, (a) maximum, (b) average and (c) minimum values.

× × × Calculated points.
○ ○ ○ Experimental points.

in sequence and moves one frame every exposure. The significant feature of this method of photography is that the final image is made up of two parts which are projected separately upon the plate. One part, the marks of degrees and minutes, repeats for each exposure and is photographed from a graticule rigidly fixed to the frame of the machine. The other part, the numbers, does not need to be positioned so very precisely, so that projection from a movable film strip is quite adequate.

A similar technique is applicable to the photographic production of binary pattern scales of the type shown in Fig. 4. A brief specification of the scale is as follows:—

Outside diameter of glass	12 in
Outside diameter of scale	11 in
Width of ring (2 half-rings)	0·01 in
Number of rings (digits)	14
Angular accuracy of position of elements of outside ring	±15 sec
Angular accuracy of other rings relative to outside ring	±1/16 length of element

It will be noted that there is considerable tolerance in the angular positioning of the elements of the inner rings, and it is convenient to divide the scale into two parts, e.g. the nine inner rings and the five outer ones. The tenth to the fourteenth rings form a section of pattern which repeats 512 times round the circumference and which may therefore be projected from a rigidly mounted graticule. The 512 sections of pattern required to complete the inner nine rings do not repeat and may be projected from consecutive frames of a film strip. Small errors in the positioning of the film are easily absorbed in the permitted tolerances.

An enlarged drawing, with rectangular sides, of the repeating section of the pattern was made by hand. An initial reduction with an ordinary stand camera provided a negative intermediate in size and about six times the size finally required. The second reduction was performed on an optical bench where special attention was given to obtaining both the final overall dimensions required (0·27 × 0·2 in) and the correct inclination (360°/512) of the sides of the pattern which corresponds to the radial direction. Final projection yielded an image on the glass disc reduced by a further factor of four. The repeating-pattern graticule is illustrated in Fig. 9(a). The variable-pattern film was made on Ilford "Microneg Ortho" film (16 mm), the size of each image frame being 0·27 × 0·36 in, or practically the

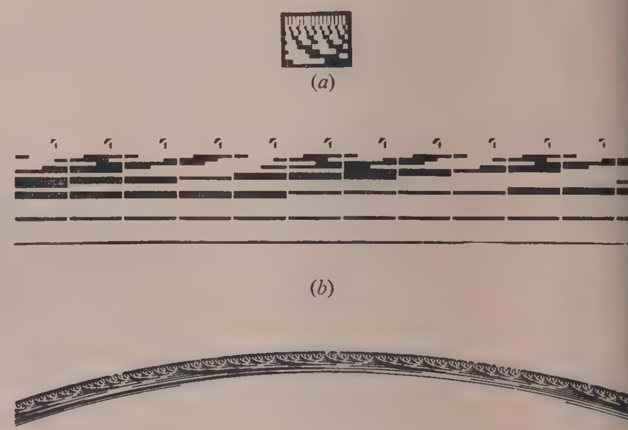


Fig. 9.—Parts of scale (actual size).

(a) Repeating pattern.
(b) Part of film strip.
(c) Part of actual scale.

total space between perforations. Each frame was photographed from a specially built pattern generator using a G45 RAF camera gun modified for taking single shots. A section of the film is shown in Fig. 9(b). The pattern generator consisted of a number of black and white slides which could be positioned by hand, together with a mask which could be superimposed on the slide for ring 8. The slides and mask had to be correctly set for each successive frame. Fig. 10 indicates the arrangement,

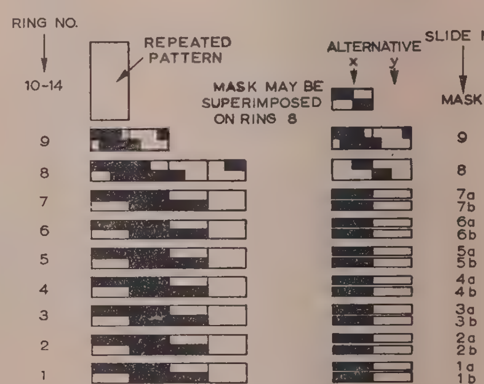


Fig. 10.—Pattern-generator slides.

in which the setting of each of the 17 variables was reduced to two alternatives. The problem of ensuring the correct sequencing of slide movements was dealt with by fitting indicating lamps behind the slides and wiring to three rotary selector switches connected to illuminate the lamps in the correct order. The final production of a complete circular pattern free from blemishes required considerable skill and patience. The double projector is illustrated diagrammatically in Fig. 11. A large

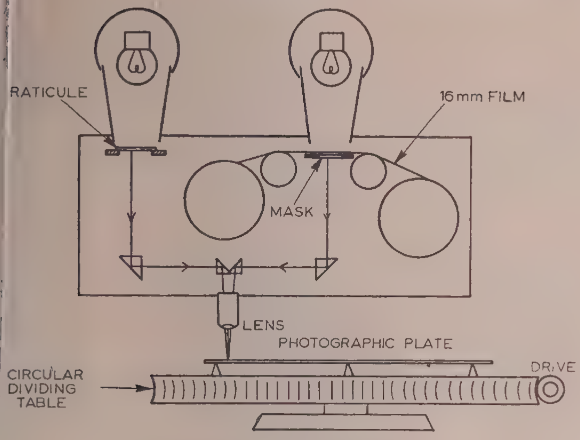


Fig. 11.—Double projector.

umber of adjustments were necessary to ensure satisfactory esults. The more important of these were as follows:—

(*a*) Focus of the graticule and film images.
(*b*) Correct orientation of the images relative to the circular track.
(*c*) Correct positioning of the images to have contact but no overlap.
(*d*) Correct sizes of the images to fit an arc of 360°/512 on a diameter of 11 in.
(*e*) Even illumination of the images.
(*f*) Levelling of the photographic plate on the work table of the dividing engine.
(*g*) Centring of the photographic plate.

ther refinements which proved necessary were a jet of filtered r to remove dust from the plate during exposure, voltage-abilizing equipment for the lamps, and automatic timing and equencing equipment. Kodak Maximum Resolution plates 5 × 12 in were used from which discs just under 12 in diameter ere cut before the exposures were made. Scales made in this ay have proved to be very satisfactory for use in experimental ata-conversion equipment. An attempt has been made to eproduce part of a scale actual size in Fig. 9(*c*).

(5.2) Scanning Equipment

Since the total width of the scale is only 0·14 in an optical ystem must be interposed between it and the cathode-ray tube. he reduction is carried out in two stages, the first of which educes 4·17 times and throws an image of the scanning spot pon a rigidly mounted slit. The slit is incorporated in order to efine the width of the spot and to ensure that the image does not ander tangentially. It also permits a small tolerance in the osition of the spot. The second stage reduces 3·36 times and mploys a 2 in microscope objective to produce the final image n the scale. Light passing through is collected by a photocell ithout the need for any further optical components. In order o produce a rigid and moderately compact structure the optical ystem is folded up using three right-angle prisms. Adjustments re provided for focusing the final image, and for rotating the athode-ray tube through a small angle about its own axis.

(5.3) Future Developments in the Technique of Scale Reading

No attempt has been made to produce an engineered version of the equipment described. This is a problem which has been considered to some extent by Elliott,[6] who describes an ingenious method of mounting the cathode-ray tube so that vibration does not cause displacement of the scanning spot. Lippel[8] approaches the subject from a slightly different point of view. A single-flash tube illuminates the whole of a radial slit at each reading instant. A separate photoelectric cell, amplifier and storage cell are provided for each digit. The assembly of photocells is rather bulky and the associated equipment rather large, but attempts are being made to use germanium photocells for this purpose, and should result in a much more compact unit.

Some of the alternatives to optical reading may be worth considering in relation to future development. For example, techniques are available for the accurate positioning of digits stored on the circumference of a magnetic drum. At present they are read by virtue of their movement past a suitable head. Second-harmonic modulators have been designed, however, by which very small static magnetic fields may be detected, and it might be possible to design static-reading heads operating upon this principle provided that the digit frequency required is not too high.

(6) ACKNOWLEDGMENTS

Acknowledgment is made to the Controller of H.M. Stationery Office for permission to publish the paper. The work was carried out at the Signals Research and Development Establishment, Ministry of Supply, and the author wishes to thank the Chief Superintendent for permission to use it as material in a thesis submitted for the degree of Doctor of Philosophy at London University. Acknowledgments are also due to Messrs. Hilger and Watts, Ltd., who produced the high-accuracy scales and optical scanning equipment, and particularly to Mr. A. Longmaid and Mr. C. B. Rhodes of their staff.

(7) REFERENCES

(1) Whitehead, E. D., and Walsh, J.: "Radio Telemetering," *Proceedings I.E.E.*, Paper No. 1389 R, March, 1953 (**100,** Part III, p. 45).
(2) Bell, J.: "Data-Transmission Systems," *Journal I.E.E.*, 1947, **94,** Part IIA, p. 222.
(3) Barker, R. H.: "A Servo System for Digital Data Transmission" (see page 52).
(4) Oxford, A. J.: "Pulse-Code Modulation Systems," *Proceedings of the Institute of Radio Engineers*, 1953, **41,** p. 859.
(5) Allard, L. S., and Hill, R. T.: "Switch and Storage Tubes," *Wireless Engineer*, 1951, **28,** p. 187.
(6) Elliott, W. S., Robbins, R. C., and Evans, D. S.: "Remote Postion Control and Indication by Digital Means," *Proceedings I.E.E.*, Paper No. 1897 M, November, 1955 (**103** B).
(7) Barker, R. H.: "Improvements in or Relating to Apparatus for the Presentation of Data in Digital Form," British Patent No. 22215/48.
(8) Lippel, B.: "A High-Precision Analog-to-Digital Converter," *Proceedings of the National Electronics Conference*, 1951, 7 (Chicago).

A SERVO SYSTEM FOR DIGITAL DATA TRANSMISSION

By R. H. BARKER, B.Sc., Ph.D., Associate Member.

(*The paper was first received* 21*st December*, 1954, *and in revised form* 7*th June*, 1955.)

SUMMARY

Certain special features must be taken into account in the design of a servo system which is to operate satisfactorily with digital data. In particular, the data are quantized in amplitude, thereby introducing non-linear effects; they constitute a series of samples instead of a continuous function, and there may be significant delays due to low-speed transmission circuits or digital equipment or both. All of these contribute in various ways towards reducing the stability of the system.

The method of synthesis takes full account of the sampling and delay features and enables a degree of prediction to be incorporated which ensures that the regenerated data do not lag on the original under steady-state conditions. Effects especially attributable to amplitude quantization are then studied qualitatively. Since stability diminishes and the sensitivity to noise and instrumental errors increases as the time of prediction is increased, it is essential that the delay involved in the transmission of each sample should be minimized. Furthermore, the sampling servo system is less efficient than a continuous one as a smoothing device, and as much smoothing as possible should be applied at the sending end before sampling.

LIST OF SYMBOLS

t = Time.
τ = Sampling period.
k = Integer identifying the instant of sampling.
p = Operator of the Laplace transform.
m = Fraction of τ by which sampling instants are effectively advanced.
$1 - m$ = Time delay as a fraction of τ, omitting whole numbers.
$\lambda\tau$ = Total time delay.
z = Operator of the sequence transforms.
w_k = Weighting sequence.
$W(z)$ = Pulse transfer function round complete feedback loop.
$Y(z)$ = Overall pulse transfer function of servo system.
$H(z)$ = First part of operational instruction of controller. (See Fig. 3.)
$\phi(p)$ = Second part of operational instruction of controller.
$v(p)$ = Transfer function of motor amplifier.
$J(z, m)$ = Pulse transfer function corresponding to $\phi(p)v(p)\varepsilon^{-p\lambda\tau}$. (See Fig. 3.)
$P(z, m)$ = Polynomial in z and m.
$Q(z)$ = Polynomial in z.
T = Time constant of velodyne.
α = Reciprocal of an integrator time-constant.
d = A constant; $\varepsilon^{-T/\tau}$ in Section 5.3.2.
a = Smoothing constant.
A, B = System constants.
g = Gain constant.
N = R.M.S. value of noise.
M^2 = Noise-power gain.

(1) INTRODUCTION

Accurate data may best be transmitted over long-distance communication circuits by using some form of digital representation, such as pulse-code modulation. The main problems concerned with such a data-transmission system are:

(*a*) the construction of suitable transducers;

(*b*) the transmission of digital data over normal communication circuits;

(*c*) the design of a suitable servo system for reconstituting the received data.

The first of these is dealt with in a companion paper[1] in which details are given of a transducer suitable for data presented in the form of shaft rotation. The second is known to be the subject of much active work, particularly in the United States, but little has been published so far. Most of the difficulties arise from the fact that, in the numerous speech circuits already in existence, very little attention has been paid to phase distortion, to which the ear is relatively insensitive but which has a profound effect on the transmission of pulses. Associated with this is the question of group synchronizing. That each digit as it arrives in time sequence must be given its correct numerical significance implies the provision of some special signal or marker to define the beginning of each group, or, alternatively, the unambiguous marking of a particular group, after which the others may be identified by counting. The author has described methods of group synchronizing elsewhere.[2]

Fig. 1 shows the flow diagram of the complete system, including

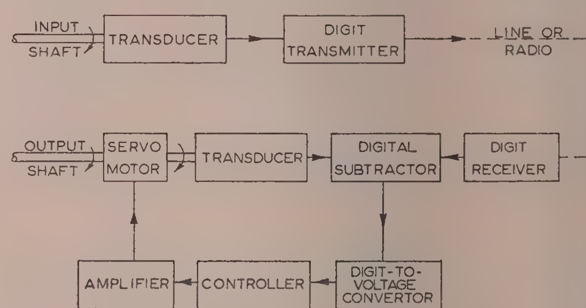


Fig. 1.—A digital data-transmission system

the essential components of the servo loop. This type of digital servo system possesses some unusual features. The error signal is not available continuously, but only after each digital subtraction has been performed. Furthermore, the sequential transmission of the digits necessitated by narrow-band circuits means that there is a delay before the error signal is available. This delay may result in a significant error unless it is made up by some form of prediction. In the particular type of coder developed for this work the digits are generated sequentially in synchronism with the transmission. Even at the receiver, therefore, there is a time delay corresponding to the duration of one number group before the measured error can be applied as a correction. That this delay is present in the feedback loop ensures that the system performs the required function of prediction. The delay also tends to reduce the stability of the

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
Dr. Barker is at the Ministry of Supply.

ystem. Finally, the error signal is not continuously variable but quantized in amplitude so that the system is non-linear, with ne non-linearities becoming more important with decreasing mounts of error. Two methods of treatment are possible. During the times that the errors are large or changing rapidly a near theory would be a good approximation, with the effect f quantizing apparent as noise. As the system approaches the eady state, on the other hand, the errors are quantized into a nall number of levels and the system may be regarded as having n-off control.

One of the earliest investigations into the theory of sampling ervo systems was conducted by Hazen,[3] who came to the conusion that the correction, applied at regular intervals, should e proportional to the measured deviation. The analytical eatment to be used here refers to such a linear system. It is ased upon a transformation closely analogous to the well-known aplace transform. In the latter, the operational quantity $\varepsilon^{-p\tau}$ presents the effect of a time delay, τ, and it is not surprising find this quantity appearing in the treatment of a system in hich successive samples enter at discrete instants τ apart. ardner and Barnes,[4] for example, use it in their discussion of e so-called "jump-function," this being a function that is only ermitted to change at regular intervals.

It was applied explicitly to sampling servo mechanisms by IacColl,[5] who introduces functions of p such as $\sum_{k=0}^{\infty} x(k\tau)\varepsilon^{-kp\tau}$, which τ is the sampling interval. The theory was further aborated by Hurewicz,[6] who suggested the operator z to place $\varepsilon^{p\tau}$. Also he introduced the important concept of the ulse transfer function of a filter as relating a sequence of input mples to the sequence of samples generated by it. He devoted pecial attention to the case in which the sampling interval is nall but not quite small enough for the system to be treated continuous.

Lawden[7] was the first to draw attention to the analogy between e summation sign in the above expressions and the integral gn in the Laplace transform. The operation was symbolized $x_k \supset X(z)$, $X(z)$ being called the sequence transform of x_k nd defined as $X(z) = \sum_{k=0}^{\infty} x_k z^{-k}$.

The basic methods of deriving and using sequence transforms ive been explained in a paper[8] which provides the necessary ckground of mathematics required for the system synthesis be detailed later and which also includes a useful table of quence transforms. This technique for dealing with sampling stems is now becoming fairly well known, and has been used various authors.[9,10,11,12]

(2) A METHOD OF SYSTEM SYNTHESIS

A system characterized by a transfer function, $w(p)$, will ossess a unit impulse response, $w(t)$, which is the inverse Laplace ansform of this and which, if sampled at successive instants , $(k = 0, 1, 2, \ldots)$, will provide the so-called weighting quence, w_k. The sequence transform of w_k is $W(z) = \sum_{k=0}^{\infty} w_k z^{-k}$, d is known as the pulse transfer function (p.t.f.) of the system. ne variable z is used in sampling-system analysis in a manner actly analogous to that in which p is used in the analysis of ntinuous systems. All the important equations in servo eory have their analogies and in many cases (though not all) ay be obtained by merely changing the variable.

An important limitation—and one easily overlooked—is that sults depending upon an equation in z refer to the sampling stants only. They provide no information as to what may be ppening between these instants, and under certain conditions a completely false impression of stability may be obtained. To overcome this difficulty a further variable, m, is introduced into the sequence transform, which is now defined as

$$W(z, m) = \sum_{k=1}^{\infty} w_k(m) z^{-k} \quad . \quad . \quad . \quad . \quad (1)$$

The weighting sequence now used is the sequence obtained by substituting $t = (k - 1 + m)\tau$ in the impulse response function, $w(t)$. The instants of sampling have effectively been advanced by an amount $m\tau$ corresponding to a system time delay of $(1 - m)\tau$. Provided that $w(t)$ is a continuous function with no jumps, the new sequence transform approaches $W(z)$ as m approaches 1, and $z^{-1}W(z)$ as m approaches 0.

Fig. 2 shows the system to be studied. The error unit embodies

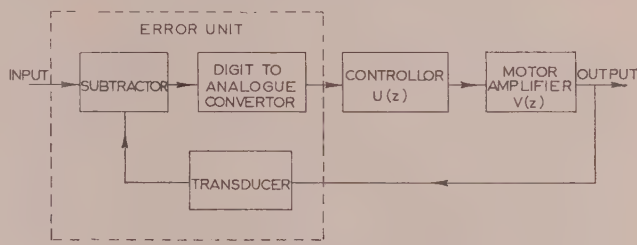


Fig. 2.—The basic digital servo system.

both the analogue-digital transducer, which periodically expresses the angular position of the output shaft as a number in binary-code form, and the digital subtractor, which finds the difference between this number and the incoming one. As explained previously, these operations cannot be conducted without a certain amount of delay, which will be denoted by $\lambda\tau$. The characteristics of the servo motor-amplifier combination depend upon the particular application. They are assumed to be known and invariant, so that the problem is to synthesize a suitable controller.

The first decision to be taken concerns performance criteria. A statistical treatment to minimize the mean square error is possible if sufficient data are available as to the sequences to be followed, but it involves formidable complications in all except the simplest cases. It is easier and, in the present state of the art, probably just as useful to assess the performance in terms of the response to specific driving functions such as a step function, a steady velocity or acceleration, a sinusoidal input of various frequencies or a random noise. Any or all such tests may be applied and, since an improvement in one respect is accompanied by deterioration in another, the final outcome must necessarily be a compromise. Such overriding factors as the demand for zero velocity-lag characteristics must of course take precedence.

The following sequence transforms will be used:—

$U(z)$ = Pulse transfer function of the controller.
$V(z)$ = Pulse transfer function of the motor-amplifier combination.
$W(z)$ = Pulse transfer function round the complete feedback loop.
$Y(z)$ = Overall pulse transfer function of the complete system.

It can be shown that

$$Y(z) = \frac{\text{p.t.f. of forward path}}{1 + W(z)} \quad . \quad . \quad . \quad . \quad (2)$$

In general the servo amplifier will be liable to overload if it is fed with a series of discontinuous pulses representing samples. Its input must in fact be reasonably smooth, and the correction

due to one error number will not be complete before the next is begun. Under these conditions the simple relationship $W(z) = U(z)V(z)$ is not valid and the system must be treated as a whole.

An artificial subdivision is, however, possible and useful. The equivalent system is shown in Fig. 3. The delay, $\lambda\tau$, in the

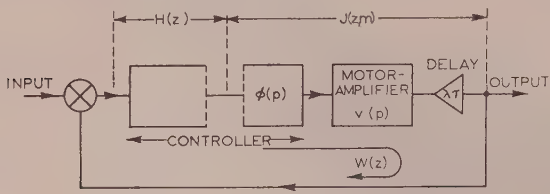


Fig. 3.—A system equivalent to that of Fig. 2.

system is now shown explicitly in association with the motor-amplifier but the transducer has been omitted since the amplitude quantization, being a non-linear phenomenon, is not yet to be taken into account. The controller is divided into two parts, the first of which is characterized by its pulse transfer function, $H(z)$, and merely modifies, in some way yet to be determined, the sequence of correction samples supplied to it. The modified sequence is smoothed by the second part, characterized by its transfer function $\phi(p)$, which is selected to provide a continuous signal suitable for driving the servo amplifier which follows it. It must be emphasized that this functional subdivision of the controller is unlikely to correspond to any physical separation of its components. The composite expression $H(z)*\phi(p)$ has been called the "operational instruction" of the controller.

That part of the system shown to the right of the dividing line has the transfer function $\phi(p)v(p)\varepsilon^{-p\lambda\tau}$ and a corresponding pulse transfer function which will be denoted by $J(z, m)$, the m being included because the delay may not be exactly equal to one or more sampling intervals. The equation $W(z, m) = H(z)J(z, m)$ is now valid. The law of motion of the output shaft between one sampling instant and the next depends upon the product $\phi(p)v(p)$. If this equals p^{-2} the motion will be linear with time; if it equals p^{-3} it will be quadratic. The factor $(1 + pT)^{-1}$ would introduce an exponential component and so on, as is well known from Laplace transform theory. To prevent the mathematics from becoming too unwieldy this expression must be kept simple and all second-order effects such as field build-up time should be discarded, at least in the initial attempt at synthesis.

From a knowledge of $v(p)$ and of the performance requirements it should be possible to specify $\phi(p)$. For example, if the motor amplifier has simple velodyne characteristics, its transfer function would be $1/p(1 + pT)$ where the time-constant, T, is probably smaller than the sampling interval. To avoid sudden changes in velocity with successive corrections it would then merely be necessary to let $\phi(p) = p^{-1}$.

The next step to be taken involves the overall performance of the system and is directed towards the determination of a suitable $W(z)$ or $W(z, m)$. One may go a long way in this direction by taking into account all the overriding factors. Use is also made of the fact that $W(z, m)$ may be expressed as the ratio of two polynomials, $P(z, m)/Q(z)$, of which it may be noted only the numerator is a function of m.

(2.1) Overriding Factors

(2.1.1) Physical Realizability.

The order of P must be at least one less than that of Q. If this were not so the series obtained by expanding in inverse powers of z, i.e. $a_0 + a_1z^{-1} + a_2z^{-2} + a_3z^{-3} + \ldots$ etc., would contain the constant term a_0 not equal to zero. A unit input at $t = k\tau$ would invoke an output of magnitude a_0 at the same instant, implying an infinitely high frequency response.

(2.1.2) Poles at $z = 1$.

For the servo system to have zero static error the function $W(z)$ or $W(z, m)$ must possess at least a simple pole at $z = 1$. A second-order pole at $z = 1$ would provide a zero velocity-lag characteristic and a third-order pole a zero acceleration-lag characteristic.

(2.1.3) Cancellation of Poles and Zeros.

The characteristic equation of the system is $Q(z) + P(z, m) = 0$. The system will be unstable if any root lies on or outside the unit circle $|z| = 1$. One may be tempted to arrange by adjustment of parameters for the cancellation of a zero by a pole so as to eliminate a root which would otherwise lie outside the unit circle. Such a procedure is not valid unless the cancellation holds good for all values of m from 0 to 1. This point cannot be emphasized too strongly, particularly since it is simpler to deal first with the special case of no delay, and the instability would then be revealed only when the behaviour between sampling instants is investigated.

(2.1.4) Design Constants.

The suggested method of synthesizing a system with suitable stability is to match the characteristic equation to one known to be satisfactory. Lawden[13] has used equations of the form $(z - a)^n = 0$, though when n is a small number a departure from this may be desirable. Whiteley[14] gives examples relevant to continuous systems and the coefficients he lists may well be suitable for sampling systems. The procedure is therefore to arrange for $P(z, m)$ and $Q(z)$ to include between them a number of constants (adjustable in the design stage) equal to the order of the characteristic equation. Since two additional constants may be included at the expense of increasing the order by one it is always possible to do this.

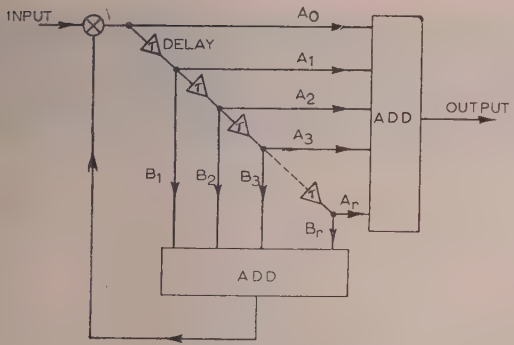
The simplest expression which meets all the above requirements is the expression to be used for W. Dividing it by $J(z)$ gives $H(z)$, the first part of the operational instruction.

(2.2) The Operational Instruction

It now remains to consider how standard electrical or electro-mechanical components may be assembled into a system possessing the required operational instruction $H(z)*\phi(p)$. The p-part, fortunately, is usually fairly simple. It must take into account the properties of the digital-analogue convertor included in the error unit. In particular, this convertor may perform the function of a clamp for which the operational instruction is $(1 - z^{-1})*p^{-1}$. Other simple functions of p may be obtained by synthesizing electrical networks in the usual way, and may or may not lead to the inclusion of further terms in the z-part. When this has been completed there is left an expression in z which is required to make up the rest of the operational instruction. A general form for such an expression is

$$\frac{A_0 + A_1z^{-1} + A_2z^{-2} + \ldots + A_rz^{-r}}{1 \;\; + B_1z^{-1} + B_2z^{-2} + \ldots + B_rz^{-r}}$$

and one way of constructing its physical counterpart with the aid of r delay elements (each equal to τ) is illustrated in Fig. 4. The output is obtained by the addition of delayed components proportional to the coefficients in the numerator. The correct denominator is obtained by negative feedback of delayed components proportional to its coefficients. An alternative is to split up the expression by partial fractions into a number of



g. 4.—A general form of construction for a unit whose p.t.f. is the ratio of two polynomials.

nple or quadratic terms and to construct the physically corre- onding exponential or oscillatory units. In the examples hich follow, however, neither of these two courses has been ırsued. Instead, use is made of the equipment necessary to ovide the p-part of the operational instruction, and the extra lays necessary to yield the correct z-part have been added her in the forward path or in a subsidiary feedback loop or both.

(3) ANALOGUE SYSTEMS

The two fundamental features which distinguish a digital rvo system from one operating on continuous data are:

(a) The data are sampled at regular instants in time.
(b) The data are quantized in amplitude.

Furthermore, with the type of data transducers described in eference 1, which were used in some of the experiments to be scribed, there are also errors of a particular type when the tum shaft is turning rapidly. These differences have as far possible been treated separately, and in this Section no count is taken of amplitude quantization and errors of position ading due to rapid movement. An analogue system has been nstructed to simulate accurately in other respects the simplest pe of digital system. The error unit, for example, includes the me delay (in this case equal to one data period, τ) as is inherent the digital equipment.

(3.1) Synthesis of a Simple Analogue System

A simple example illustrating the method of synthesis outlined ove and applicable later to a digital system may be helpful here. g. 3 is the starting-point and the following assumptions are de with regard to it:—

a) The servo motor and amplifier are constructed as a simple odyne. The rate of rotation is proportional to the applied tage so that the transfer function is $v(p) = Ap^{-1}$.

b) The transducer performs the function of sampling and ludes a delay, τ, the effect of which is to multiply the p.t.f. z^{-1}.

c) There must be zero static error. Hence $W(z)$ must include factor $(z - 1)$ in its denominator.

d) There shall be no sudden changes in output velocity. The of motion is to be quadratic between sampling instants. nce $\phi(p) = p^{-2}$. Therefore $\phi(p)v(p) = Ap^{-3}$ and $J(z)$ $\tfrac{1}{2}Az(z+1)/(z-1)^3$, the transform being obtained directly m the Table given in Reference 1.

The loop p.t.f. may now be written as

$$W(z) = \frac{\tfrac{1}{2}AH(z)(z+1)}{(z-1)^3} \quad . \quad . \quad . \quad . \quad . \quad (3)$$

and contains the required factor $(z - 1)$ in the denominator. The remaining requirement to be fulfilled by $H(z)$ is that it shall include design constants so that the characteristic equation may be identified with one known to be suitable. The equation is so far of third order and already includes the gain constant, A. The simplest expression for $H(z)$ which introduces two further constants without increasing the order is

$$H(z) = \frac{(z-1)^2}{z^2 + B_1 z + B_2} \quad . \quad . \quad . \quad . \quad . \quad (4)$$

The operational instruction of the controller is therefore to be

$$\frac{(z-1)^2}{z^2 + B_1 z + B_2} * p^{-2}$$

Its pulse transfer function is obtained by replacing p^{-2} with its sequence transform and is $z/(z^2 + B_1 z + B_2)$. Since the order of the numerator is one less than that of the denominator there is no difficulty about physical realizability. The characteristic equation is

$$z^3 + (B_1 - 1)z^2 + (B_2 - B_1 + \tfrac{1}{2}A)z + (\tfrac{1}{2}A - B_2) = 0$$

The simplest third-order equation with which this can be identified is $z^3 = 0$. Choice of an equation of the form $z^n = 0$ is known to result in the most rapid recovery from a transient disturbance. The required conditions are that $A = B_1 = 1$ and $B_2 = \tfrac{1}{2}$, and, if these are fulfilled,

$$W(z) = \frac{z + 1}{2z^3 - z - 1}$$

and the overall p.t.f. of the system is

$$Y(z) = \frac{W(z)}{1 + W(z)} = \tfrac{1}{2}z^{-2} + \tfrac{1}{2}z^{-3} \quad . \quad . \quad . \quad (5)$$

(3.2) The Controller

The p-part of the operational instruction is p^{-2} and can be introduced physically either by a two-stage integrator or by a clamp followed by a single stage of integration. The latter happens to be easier to apply with the experimental techniques available, and has been used. The clamp, however, has the factor $(1 - z^{-1})$ which must be allowed for. Two subsidiary feedback loops will be needed to obtain a quadratic expression in the denominator, so that a likely arrangement for the controller is that of Fig. 5. The integrator in the forward path

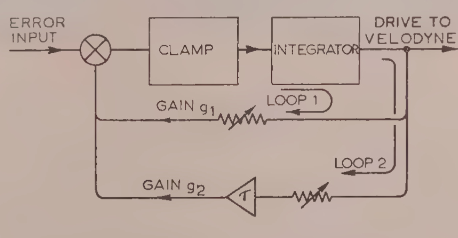


Fig. 5.—Controller for the simple digital system.

effectively introduces the required delay in loop 1. The p.t.f. of this arrangement is obtained as follows:—

P.T.F. of forward path $= U_1(z) = (1 - z^{-1})\dfrac{z}{(z-1)^2} = \dfrac{1}{z-1}$

P.T.F. allowing for loop 1 with gain g_1 $= U_2(z) = \dfrac{U_1(z)}{1 + g_1 U_1(z)} = \dfrac{1}{z + g_1 - 1}$

P.T.F. allowing also for loop 2 with gain g_2
$$= U(z) = \frac{U_2(z)}{1 + g_2 z^{-1} U_2(z)}$$
$$= \frac{z}{z^2 + (g_1 - 1)z + g_2}$$

The last expression is exactly that required if $g_1 = 2$ and $g_2 = \frac{1}{2}$.

(3.3) Experimental Verification

In the experiments to verify this theory the input circuit of the clamp was an amplifier with a large amount of negative feedback. The loop gains g_1 and g_2 were adjustable by means of decade resistance boxes, as was the main-loop gain, A. The gains could be accurately set to the required values by a calibration process which is described briefly here since it throws light on the manner in which the controller stabilizes the system.

(3.3.1) Calibration of Loop 1 Gain.

Loop 2 and the main loop are opened so that $g_2 = 0$. The p.t.f. of the controller reduces to $U_2(z) = 1/(z + 1)$ so that the correct gain-control setting ($g_1 = 2$) is such that any oscillations set up are just sustained with constant amplitude. Fig. 6 shows the

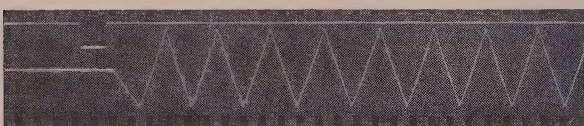


Fig. 6.—Calibration of loop-1 gain.
Upper trace: Unit sample of error.
Lower trace: Resulting constant-amplitude oscillation.

response of the unit in this condition to a single unit sample (upper trace). The sampling instants have been marked by brightening pulses. Note the one data-period delay between the instant at which the input sample is taken and that at which the controller output begins to change.

(3.3.2) Calibration of Loop 2 Gain

Loop 1 and the main loop are opened so that $g_1 = 0$. The stability is critical when the p.t.f. is $z/(z^2 - z + 1)$, that is when $g_2 = 1$, twice the value of the normal setting. The period of oscillation is 6τ as shown in Fig. 7. A useful check is available

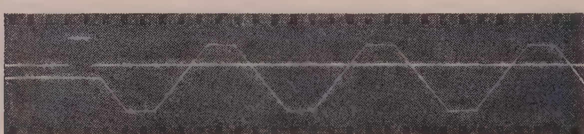


Fig. 7.—Calibration of loop-2 gain.

at this stage in that an oscillation of constant amplitude but of period 3τ is also obtained when both subsidiary loops are closed. This corresponds to the p.t.f. $z/(z^2 + z + 1)$. The only change now required is to halve accurately the gain of the second loop, after which the controller is correctly set up. The weighting sequence may be obtained as the coefficients of inverse powers of z in the expansion of the p.t.f. It is

$$0,\ 1,\ -1,\ \tfrac{1}{2},\ 0,\ -\tfrac{1}{4},\ \tfrac{1}{4},\ -\tfrac{1}{8},\ 0,\ \tfrac{1}{16},\ -\tfrac{1}{16},\ \tfrac{1}{32},\ 0 \ldots, \text{etc.}$$

Fig. 8 shows the response of the controller to unit sample, and the recorded output values at the sampling instants (brightened) correspond accurately to this sequence.

(3.3.3) Calibration of the Main-Loop Gain.

The gain of the main loop may be calibrated after the controller has been correctly set up by increasing A until the whole system

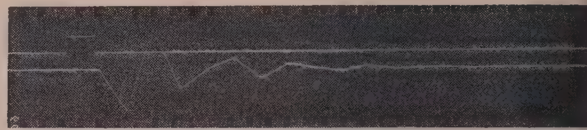


Fig. 8.—Normal response of controller to unit sample.

is just capable of maintaining oscillations. The characteristic equation is then

$$2z^3 + (A - 1)z + (A - 1) = 0$$

and this must have one pair of complex conjugate roots lying on the unit circle $|z| = 1$, and one real root within it. It can therefore be written also in the form

$$(z + x)(z^2 - xz + 1) = 0$$

in which, by equating coefficients,

$$2x = 2 - 2x^2 = A - 1 \quad .$$

By solving the quadratic in x the critical value for the main-loop gain may be obtained, namely $A = \sqrt{5}$.

Merely for the purpose of calibration it is unnecessary to proceed further, but it is of interest to note that the p.t.f. of the system set up in this condition is

$$Y(z) = \frac{0 \cdot 242}{z + 0 \cdot 618} - \frac{0 \cdot 242z - 1 \cdot 417}{z^2 - 0 \cdot 618z + 1} \quad .$$

The first term corresponds to a decaying oscillation of period 2τ; the second corresponds to an undamped oscillation of period 5τ for which the weighting sequence is

$$w_k = -0 \cdot 242,\ 1 \cdot 267,\ 1 \cdot 026,\ -0 \cdot 633,\ -1 \cdot 417,\ -0 \cdot 242,\ 1 \cdot 267,\ \ldots, \text{etc.}$$

Fig. 9 shows both the initial transient and the sinusoidal sequence. The law of motion between the sampling instants is of course parabolic, not sinusoidal.

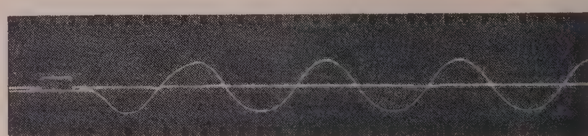


Fig. 9.—Calibration of main-loop gain; oscillation initiated by unit sample, period 5τ.

(3.3.4) Overall Performance.

The overall p.t.f. of the system properly set up is $\frac{1}{2}z^{-2} + \frac{1}{2}z^{-3}$, so that the response to a unit step input is the sequence corresponding to $(\frac{1}{2}z^{-1} + \frac{1}{2}z^{-2})/(z - 1)$, i.e. 0, 0, $\frac{1}{2}$, 1, 1, 1, 1, etc. At the end of the delay period the output moves with constant acceleration until half-way and then moves with equal deceleration until the rest position is reached. Fig. 10 shows this move-

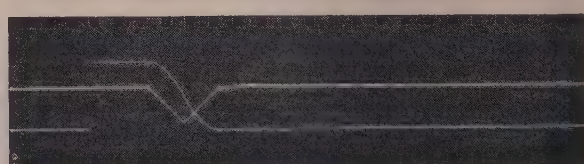


Fig. 10.—Response of servo system to unit step.
Upper trace: Controller output voltage.
Lower trace: Error signal.

:nt in terms of the error signal. The second trace with the shaped deflection is the voltage which the controller supplies the velodyne integrator.

(3.4) **Effects of Imperfections**

The relatively simple theory used in the synthesis of this stem ignored all second-order effects. In the experimental uipment neither the clamp nor the integrators were perfect. ie operational instruction of the clamp, for example, was more arly $\frac{z-1}{z+d}*p^{-1}$, with $d = 0{\cdot}02$. The transfer function of the ntroller integrator approximated to $(p+\alpha)^{-1}$, $(\alpha = 0{\cdot}05)$ d that of the velodyne integrator to $p^{-1}(1+pT)^{-1}$, with $\tau = 0{\cdot}001$. The revised theoretical treatment can be carried rough without increasing the order of the characteristic uation and with the new constants, d, α and T appearing in rrection terms. The effects of these corrections are largely mpensated for by the calibration procedure which has been opted, and an accurate expression for the p.t.f. of the controller $(z+h)/(z^2+z+\frac{1}{2})$ where $h = 0{\cdot}05$.

The calibration of the main-loop and subsidiary-loop gain ntrols was carried out with reasonable care and accuracy to ovide experimental data with which to verify the theory. ch a procedure is not of course necessary merely to set up e system to have a satisfactory performance. The main-loop in, for example, may be more than doubled before the system gins to be unstable, and the controller is little affected by gain riations of the order of $\pm 10\%$; such limits are easily held th the aid of feedback-stabilized amplifiers.

ANALOGUE SERVO SYSTEM WITH ZERO VELOCITY LAG

(4.1) **System Synthesis**

The basic assumptions to be made in commencing the syn- esis of a zero velocity-lag system are exactly the same as those Section 3.1 except that $W(z)$ must now include the factor $-1)^2$ in its denominator. The loop p.t.f. may again be itten as $W(z) = \frac{1}{2}AH(z)(z+1)/(z-1)^3$, but we shall only permitted to cancel one of the poles at $z = 1$. It is not ssible, therefore, to avoid increasing the order of the charac- istic equation to four, and three constants additional to A e needed if the characteristic equation is to be put in the desired rm. The simplest expression for $H(z)$ including three con- nts and the factor $(z-1)$ in its numerator is $H(z) =$ $-1)(z-B_3)/(z^2+B_1z+B_2)$. The operational instruction the controller is therefore $\frac{(z-1)(z-B_3)}{z^2+B_1z+B_2}*p^{-2}$ and its p.t.f. is

$$U(z) = \frac{z(z-B_3)}{(z^2+B_1z+B_2)(z-1)}$$

e loop p.t.f. is

$$W(z) = \frac{\frac{1}{2}A(z+1)(z-B_3)}{(z^2+B_1z+B_2)(z-1)^2}$$

d the characteristic equation is

$$+ (B_1-2)z^3 + (1+\tfrac{1}{2}A-2B_1+B_2)z^2$$
$$+ (\tfrac{1}{2}A - \tfrac{1}{2}AB_3 + B_1 - 2B_2)z + (B_2 - \tfrac{1}{2}AB_3) = 0 \quad . \quad (6)$$

e coefficients of this equation may be made identical with ose of any fourth-order equation known to be suitable. The uation $z^4 = 0$ would permit the most rapid recovery from a nsient disturbance, in which case the response to a unit step uld be the sequence 0, 0, $1\frac{3}{4}$, $2\frac{1}{4}$, 1, 1, 1, etc. Both the error d the summated errors become zero within three sampling periods of the application of the correction to the controller. The overshoot is very large, however, and some degree of smoothing is clearly required. The characteristic equation will therefore be identified with $(z-a)^4 = 0$, where a is a positive smoothing constant. The conditions for this identity are

$$B_1 = 2 - 4a$$
$$4B_2 = 4a^4 + (1-a)^3(5+3a) = 5 - 12a + 6a^2 + 4a^3 + a^4$$
$$B_3 = (5+3a)/(7+a)$$
$$2A = (1-a)^3(7+a)$$

A method by which the controller can be constructed becomes apparent when the expression for $U(z)$ is split into partial fractions thus

$$U(z) = \frac{A_1}{A}\,\frac{z+h}{z^2+B_1z+B_2} + \frac{A_2}{A}\,\frac{1}{z-1} \quad . \quad . \quad (7)$$

in which $$A = A_1 + A_2$$

Provision must be made for adjustment of the constants A_1, A_2 and h together with B_1 and B_2 and for a calibration procedure similar to that used in the previous example. These constants are plotted as a function of a in Fig. 11. Inspection of eqn. (7)

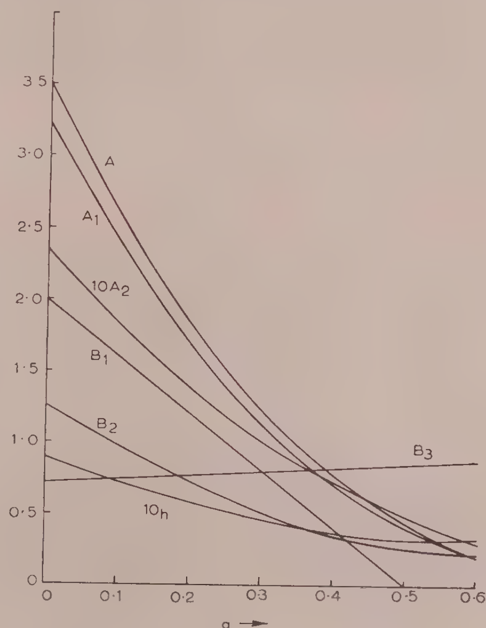


Fig. 11.—Control-constant values for zero-velocity-lag servo system.

reveals that the form of the first term of the expression for $U(z)$ is exactly that relating to the controller described in Section 3. The second term may be added by connecting an integrator in parallel with this device. The complete system is as shown in Fig. 12. It may be noted that no provision has been made for the adjustment of h. To do so would involve rather an unnecessary complication, since h varies only slowly with a, having a minimum at $a = 0{\cdot}5$ approximately. Furthermore, the experimentally determined value, $h = 0{\cdot}05$, applying to the system is about right for the amounts of smoothing likely to be used.

(4.2) **Calibration**

The resistors for determining the loop-gain constants A_1, g_1 and g_2 may be calibrated in the same way as for the previous

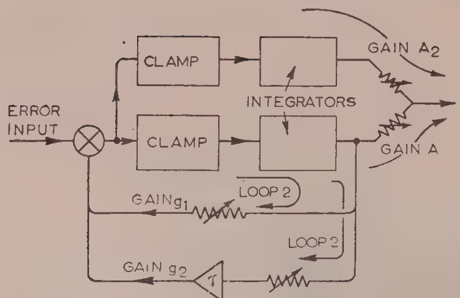


Fig. 12.—Controller for zero-velocity-lag system.

example. That part of the controller involving the subsidiary loops with gains $g_1 = 1 + B_1$ and $g_2 = B_2$ will be referred to as the stabilizer. Its p.t.f. is given by the first term of eqn. (7), from which it appears that under certain conditions the stabilizer may itself be an unstable component. In particular, the characteristic equation $z^2 + B_1 z + B_2 = 0$ has roots outside the unit circle $|z| = 1$ if $a < 0{\cdot}088$ and the oscillatory sequence corresponding to $z/(z^2 + B_1 z + B_2)$ is then of increasing amplitude.

The particular case of no smoothing ($a = 0$) has been studied to illustrate this instability. The appropriate values of the control constants are $B_1 = 2$ and $B_2 = 1{\cdot}25$. The weighting sequence is

$$0,\ 1,\ -2,\ 2{\cdot}75,\ -3,\ 2{\cdot}56,\ -1{\cdot}375,\ \text{etc.}$$

which becomes, when $(z + h)$ is used as the numerator,

$$0,\ 1,\ -1{\cdot}95,\ 2{\cdot}65,\ -2{\cdot}86,\ 2{\cdot}41,\ -1{\cdot}25,\ \text{etc.}$$

It is plotted in Fig. 13, which shows that the sequence consists of samples taken alternately from one or other of the exponentially growing sine-waves. Fig. 14 is an oscillogram obtained with the stabilizer set up in this way, and measurements made upon it agree accurately with the calculated weighting sequence.

The integrator loop gain, A_2, may be calibrated by first setting up the equipment in some known stable condition with the loop opened and then increasing A_2 until oscillations are just sustained. This critical value of A_2 may then be calculated from the measured angular frequency, β, of the oscillation without the need for finding the roots of the characteristic equation. The method is a general one and depends upon the fact that, in the critical condition, the term $(z^2 - 2z\cos\beta\tau + 1)$ will be a factor of the characteristic equation, and therefore $\varepsilon^{j\beta\tau}$ will be a root. Substitution of $z = \varepsilon^{j\beta\tau}$ gives two equations, one relating to the real part and the other to the imaginary part. Both are linear in A_2

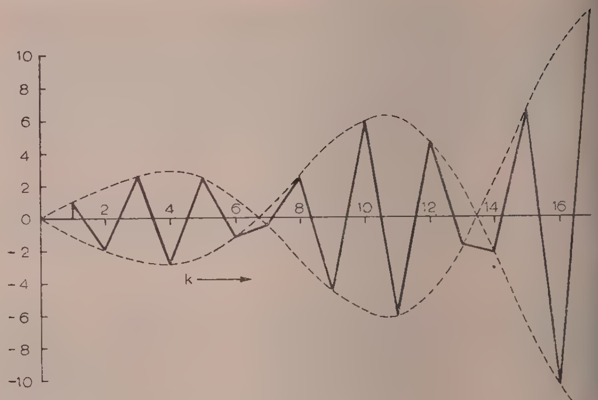


Fig. 13.—Weighting sequence of stabilizer when $a = 0$.

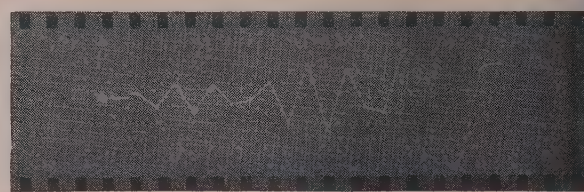


Fig. 14.—Oscillogram corresponding to Fig. 13.

and straightforward to solve. If β is known accurately, both lead to the same solution. In practice the value of β determined experimentally will have to be slightly modified by trial until such agreement is reached.

The initial condition in which the system is set up is not important except that it must be a stable one. That corresponding to the servo system described in Section 3, for example, is suitable, the values of the constants being $A_1 = 1$, $B_1 = 1$, $B_2 = \frac{1}{2}$, and $h = 0{\cdot}05$ (assumed). Experimentally it was found that $\beta\tau = 32{\cdot}7°$. This had to be corrected to $33{\cdot}15°$ for the two solutions mentioned above to agree, with the corresponding value for the critical gain, $A_2 = 0{\cdot}256$.

Table 1 summarizes the various setting-up and calibrating procedures that have been described, both in this Section and in Section 3. Only one operation is of course necessary for each of the constants involved; the remainder have been included for illustrative purposes.

Table 1

SUMMARY OF SETTING-UP PROCEDURES

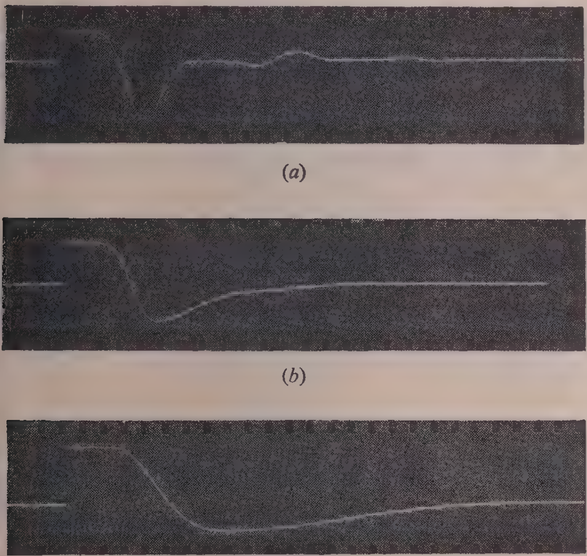
Characteristic equation	A_1	A_2	g_1	g_2	Remarks	Figure
$z + 1 = 0$	0	0	2	0	Loop 1 oscillates	6
$z^2 - z + 1 = 0$	0	0	0	1	Loop 2 oscillates	7
$z^2 + z + 1 = 0$	0	0	2	1	Combined feedback for check purposes	—
$z^2 + z + \frac{1}{2} = 0$	0	0	2	0·5	Normal setting of controller. (Section 3)	8
$2z^3 + (\sqrt{5} - 1)z + \sqrt{5} - 1 = 0$	$\sqrt{5}$	0	2	0·5	Main loop oscillates	9
$z^3 = 0$	1	0	2	0·5	Normal setting of servo. (Section 3)	10
$z^2 + 2z + 1{\cdot}25 = 0$	0	0	3	1·25	Stabilizer oscillates	14
	1	0·256	2	0·5	Critical gain in integrator loop	—
$z^4 = 0$	3·265	0·235	3	1·25	Normal setting of zero-velocity-lag servo without smoothing	15(*a*)
$(z - 0{\cdot}2)^4 = 0$	1·69	0·138	2·21	0·73	Normal (z.v.l.) with smoothing, $a = 0{\cdot}2$	15(*b*)
$(z - 0{\cdot}4)^4 = 0$	0·70	0·072	1·41	0·39	Normal (z.v.l.) with smoothing, $a = 0{\cdot}4$	15(*c*)

(4.3) Performance

ie overall p.t.f. of the system, expressed in terms of a is

$$Y(z) = \frac{(1-a)^3[(7+a)z^2 + (2-2a)z - 5 - 3a]}{4(z-a)^4} \quad . \quad (9)$$

ie fact that the value of h has not been correctly set up has ot been allowed for, but discrepancies due to this cause should ot be large. In the special case $a = 0$, the p.t.f. reduces simply $1\frac{3}{4}z^{-2} + \frac{1}{2}z^{-3} - 1\frac{1}{4}z^{-4}$, and the unit-step response is the quence 0, 0, $1\frac{3}{4}$, $2\frac{1}{4}$, 1, 1, etc. Fig. 15 shows the step response



(a)

(b)

(c)

Fig. 15.—Servo response to unit step (error signal displayed).

(a) No smoothing.
(b) $a = 0{\cdot}2$.
(c) $a = 0{\cdot}4$.

ith different values of a, the error signal being displayed to emonstrate the amount of delay involved before any movement kes place. The most significant feature of the trace for $a = 0$ the marked deviation from the zero line after the end of the itial transient. In amplitude it amounts to approximately 10% the main peak. The timing and general form of this secondary sturbance were more or less independent of small changes in e adjustment of the system constants, and although no precise planation has been found it is thought to be attributable to accurate integration by the velodyne. In assessing these perimental results it must be borne in mind that they are tended to serve as verification of the theoretical treatment, d a system such as this, with no smoothing, is too near to the argin of instability to be regarded as a practical proposition.

It is perhaps desirable in discussing these discrepancies to esent at least a crude physical explanation of the functioning the stabilizer. The main loop and loop 2 may be looked upon in parallel with respect to the integrator unit of the stabilizer, d the total delay round each of these paths is 2τ. The main op involves a phase reversal but the other does not, so that gnals arriving back at the input of the integrator tend to cancel ch other. The difference should be such that, when integrated, just suffices to drive the servo motor in the required manner. will be realized that when there is little or no smoothing the ansient response is very violent and the subsequent drive to e motor depends upon the integrated difference of two much larger quantities. Quite small departures from perfection on the part of the mechanical system will give rise to significant errors after a delay of 2τ. Measurements made on oscillograms (b) and (c) of Fig. 15 agree accurately with the calculated responses.

The effects of overloading have been examined. They are most serious when there is little smoothing, and can give rise to subsidiary damped oscillations.

An important factor is the margin of stability against variations

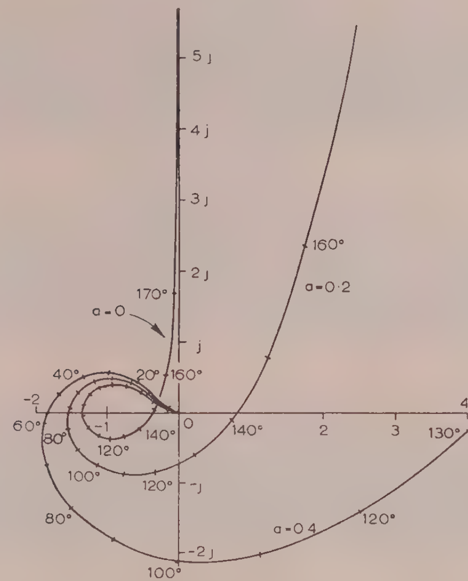


Fig. 16.—Stability diagram.

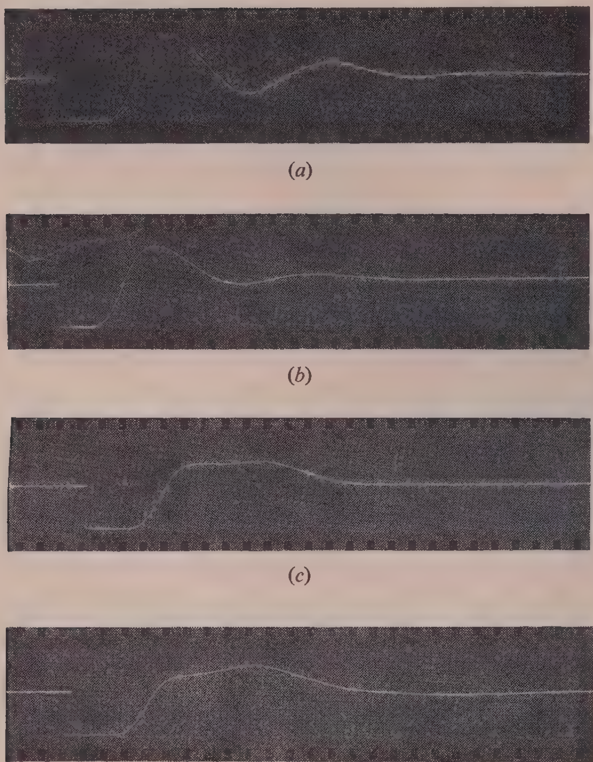
in gain. Fig. 16 is a plot of the stability diagram [real and imaginary parts of $1/W(\varepsilon^{j\omega\tau})$] for the system, from which gain margins as follows may be deduced:

$$\text{Gain margin} = 32\% \text{ if } a = 0$$
$$53\% \text{ if } a = 0{\cdot}2$$
$$82\% \text{ if } a = 0{\cdot}4$$

These figures refer to the simultaneous variation of A_1 and A_2. Fig. 17 illustrates the effect of varying the main-loop gain, A_1, only. It is apparent that moderate changes are permissible without causing instability, and that a reduction of gain is preferable to an increase. In fact oscillogram (c) reveals a response which might well be preferred to that with the so-called correct setting, for the overshoot is much reduced but the decay time is only slightly increased. Small variations of the component of integral-of-error control have little effect. A very large increase causes a low-frequency oscillation, as was demonstrated in the calibration procedure. A reduction in gain introduces a higher-frequency oscillation which rapidly dies out even when the gain is zero.

Adjustment of the stabilizer, on the other hand, is much more critical. A small error in the setting of g_1 gives rise to a lightly damped oscillation of period 2τ, and it is desirable that g_1 be accurate to within 1 or 2%. The setting of g_2 is not quite so critical and a tolerance of $\pm 5\%$ could probably be permitted.

The oscillograms obtained and the measurements made on them substantiate the general conclusion that sampling systems with delayed error may be stabilized by subsidiary feedback loops. The next step is to substitute digital for analogue control, and to investigate those features which are peculiar to digital as distinct from sampling servo systems.



(a)

(b)

(c)

(d)

Fig. 17.—Effect of varying main-loop gain (error signal response to unit step).

(a) $A_1 + 20\%$.
(b) $A_1 + 10\%$.
(c) $A_1 - 10\%$.
(d) $A_1 - 20\%$.

(5) DIGITAL SYSTEMS

The purpose of the experiments described in the preceding Sections has been to simulate those features of a digital servo system which are susceptible to linear mathematical treatment, and to examine particularly the difficulties introduced by sampling and delayed feedback.

Except in some very special circumstances, the principle of superposition does not apply to systems including amplitude-quantizing devices, and there is no suitable mathematical technique available for dealing in a general way with such problems. Two limiting cases may, however, be considered. The first of these applies when the quantum step is much smaller than the signals circulating in the system. Each quantized datum may then be regarded as comprised of that same datum before it was quantized together with a small error. Since there will usually be little or no correlation between the amount of this error in successive quantizing and sampling operations, the quantizer may be regarded as a generator of noise, the amplitude distribution of which is flat over the range of one quantum step.

The second case is when the input shaft is stationary or nearly so, for the signals then present in various parts of the system are no longer large compared to the quantum step. This applies particularly to the way in which the system comes to rest, and to the stability when in that condition.

One further source of noise or distortion is to be found in the particular type of data transducer used in the experiments and described in Reference 1. Certain dynamic scale-reading errors are introduced when the datum shaft is moving. The effect is small at low speeds, but as the speed increases the quantizing errors become biased in the direction of rotation and may in fact spread over several quantum steps. The distribution in amplitude becomes more nearly Gaussian about a mean which varies almost linearly with the velocity.

(5.1) The Error Detector

In the experimental system both input and output data are represented by binary numbers. The least significant digit comes first in time and is preceded by a frame synchronizing pulse which, in the number representation, is equivalent to the digit 1. Addition on the binary scale follows rules very similar to those of conventional arithmetic. It may be performed serially digit by digit commencing with that of least significance. At each stage there may be up to three units to be added, the addend, the addendum and the carry figure from the previous stage. The total, modulus 2, is a digit of the required answer, and if two or more of the input digits are 1's there will be a 1 to carry into the next stage. These rules are embodied in Table 2, in which A and B are digits of the numbers to be added and C is the digit carried over from the previous stage.

Table 2

Rules of Binary Addition

A	B	C	Sum	Carry
0	0	0	0	0
1	0	0	1	0
0	1	0	1	0
0	0	1	1	0
1	1	0	0	1
1	0	1	0	1
0	1	1	0	1
1	1	1	1	1

As an example, the numbers 78 and 39 are added together in binary notation.

```
A               1001110 =  78
B               0100111 =  39
C (in)          001110
                -------
Sum             1110101 = 117
                -------
Carry (out)     0001110
```

The addition is performed working as usual from right to left. A difficulty of convention arises when the numbers are represented as electrical pulses and displayed on an oscilloscope or waveform diagram. It is then usual for time to be represented as increasing from left to right. Williams, Robinson and Kilburn[15] advocate writing the arithmetic backwards to conform to the normal cathode-ray-tube displays. It seems, however, that unless the amount of data presented is large so that one becomes thoroughly familiar with the procedure there will always be some danger of confusion. Both the normal conventions are followed here. When a number is written in figures the least significant digit is to the right; when it is represented by pulses the least significant digit is to the left.

Fig. 18 shows the logical diagram of the serial binary adder used. The physical embodiment of each of the paths A, B and C is actually a two-wire circuit in which one wire or the other is raised to a positive potential according to whether the digit is 1 or 0. This enables the modulus-2 adder to consist of a simple matrix of 16 rectifiers, and the coincidence unit of a matrix of

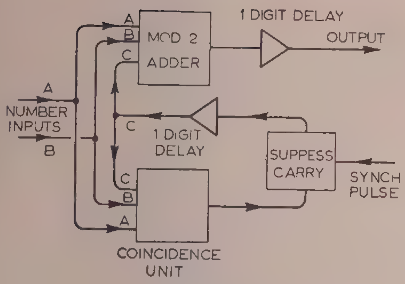


Fig. 18.—Logical diagram of binary adder.

ectifiers. Following the modulus-2 adder is a unit providing elay of one digit period. This serves to regenerate the digit ses and also rounds up to τ the total delay in deriving the or. The carry figure is delayed by one digit period and made ilable again on the two-wire circuit, C. It is necessary to vent a carry figure resulting from the addition of digits in synchronizing-pulse position from being added into the least nificant digit of the number immediately following. There is tunately a pulse available from the transducer at the correct tant to override the output from the coincidence unit, and it is d to suppress this particular carry figure.

So far only the addition of binary numbers has been described. btraction involves the representation of negative numbers. dealing with angular data for which continuous rotation is ssible the transition from 1111111111 to 0000000000 results m a change of one quantum. It would be equally satisfactory count anticlockwise from the zero position for positive numbers l clockwise for negative ones. The two binary numbers 1111111 and -1 would then be indistinguishable, which is it should be since a 10-digit system can resolve only 1 024 ferent numbers. This leads to the idea of representing negative mbers by their complement with respect to 2^n, where n is the it capacity of the device concerned. Since the addition of a nber and its complement yields 2^n, which is indistinguishable m zero, it follows that subtraction may be performed by ming the complement of one number and adding.

t is much easier electrically to form the complement with pect to $(2^n - 1)$ than to 2^n, and in this instance it merely uires the interchange of a pair of wires. The result obtained l be in error by one, but may be corrected by taking the nplement once more. This is illustrated by the following two mples, in which the frame-synchronizing pulse is also shown, arated by a comma.

70 − 67

	S	
Input number	0001000110,1	= 70
Complement	1110111001,0	
Output number	0001000011,1	= 67
Sum	1111111100,1	
Complement	0000000011,0	= 3

70 − 73

	S	
Input number	0001000110,1	= 70
Complement	1110111001,0	
Output number	0001001001,1	= 73
Sum	(1)0000000010,1	
Complement	1111111101,0	= − 3

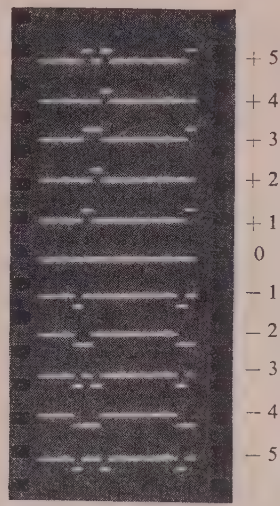


Fig. 19.—Representation of digital errors.

Fig. 19 is an oscillogram illustrating the representation of small positive and negative errors at the output of the digit delay in Fig. 18. The sampling instant is indicated by the brightening pulse. Immediately following is the most significant digit of the preceding number, then the S-digit and then the least significant and subsequent digits of the number displayed.

The voltage analogue of the error is needed to drive the servo controller. With the number representation adopted, this decoding is a simple operation. Advantage is taken of the fact that the least significant digit comes first so that time weighting may be used. Fig. 20 shows the basic circuit. The voltage across capacitance C will decay to half its initial value V in one digit period (20 ms) if the capacitance is shunted by a resistance R such that the time-constant CR equals 0·693 of the digit period.

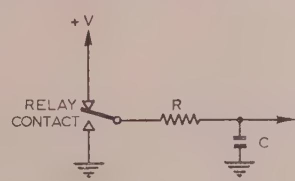


Fig. 20.—Basic decoding network.

It will decay to $\frac{1}{4}V$ in two digit periods, to $\frac{1}{8}V$ in three digit periods and so on. The same capacitor can also be used to add together the contributions of the various digits. Relay R_1 operates according to the binary signal to be decoded, being subjected to a voltage of $+V$ during digits of value 1 and down to earth during digits of value 0. Imagine the capacitor to be initially discharged. If the least significant digit is 1 and all others 0, the capacitor will charge to $\frac{1}{2}V$ during that digit and decay to $V/512$ by the end of the ninth digit. This instant is marked by the brightening pulse in Fig. 19. For an error in which the second digit is 1 and all others 0 the capacitor again charges to $\frac{1}{2}V$ but only has time to decay to $V/128$. Similarly if the error is 4 the voltage remaining at the sampling instant is $V/64$, or $4V/512$. If more than one of the digits are 1's the effects are additive and the voltage remaining at the next sampling instant is proportional to the binary number.

That an error which is larger than half a revolution is treated as negative follows from the method, already explained, of representing negative numbers. Whether the error is positive or negative depends upon the most significant digit, which now takes on the role of sign digit. Either the voltage stored on the capacitor or its negative equivalent is used according to the value of this digit.

(5.2) A Simple Digital System

A direct comparison was desirable between a digital system and the analogue system set up as previously described to simulate it. The equipment was therefore arranged so that the change-over from one to the other could be made by a single switch and without any readjustment of the gain controls. Effects due to quantizing and dynamic scale-reading errors could thereby be examined separately, in the first instance for the simple system described in Section 3.

(5.2.1) Stability in the Rest Condition.

In the digital servo system, movement of the output shaft within the limits of one element of the code is not detectable by the error-sensing unit. If it is assumed that the system comes to rest within these limits it is inevitable that the drift in the velodyne amplifier or elsewhere will sooner or later cause movement until an error of one digit appears. It is a requirement that this shall be capable of resetting the system to zero, for otherwise static accuracy would be lost.

The sequence generated by the controller in response to unit step is 0, 1, 0, $\frac{1}{2}$, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{8}$, $\frac{3}{8}$, $\frac{7}{16}$, . . ., etc., approaching two-thirds as a limit. If the initial peak (1) should fail to start the motor it is unlikely that it will start subsequently. The minimum sensitivity of the velodyne is therefore determined by the condition that the motor should start reliably on a signal corresponding to 2/3 of one element, for it is then impossible for a static error to persist.

The frequency with which the output wanders to one side and is corrected depends upon the drift stability of the amplifiers used, particularly that of the velodyne. The signal level between decoder and velodyne should be kept as high as possible consistent with freedom from limiting and overloading. With no other special precautions the frequency of correction could be reduced to about three or four per minute.

(5.2.2) Noise Magnification.

The amount of noise, N, introduced by the transducers has been estimated in Reference 1 as a function of the speed of rotation. Since there is little correlation between that generated at the receiver and that generated at the transmitter, the total effective noise power at the input of the servo system will be $2N^2$. If the noise-power gain of the system is M^2 the output noise power will be $2N^2M^2$. Experimentally it is much easier to measure the noise superimposed upon the error signal than to measure the random variations of the output signal. The noise output from the error detector is compounded of $2N^2$, which would be present if the system had infinite smoothing, plus $2N^2M^2$ due to noise fed back from the output. Hence the r.m.s. noise on the error signal is $N_e = N[2(1 + M^2)]^{\frac{1}{2}}$. In this case, where $Y(z) = \frac{1}{2}z^{-2} + \frac{1}{2}z^{-3}$, the noise-power gain, M^2, equals one half; hence $N_e = \sqrt{(3)}N$. Digital records have been taken under conditions of steady rotation and enable the noise on the digital error to be measured. Some results are given in Table 3.

As might be expected, the actual noise was greater than that calculated, for in the calculation account was not taken of factors such as uneven running of the servo, particularly at low speeds, and variations in frictional losses due to imperfections in the gearing.

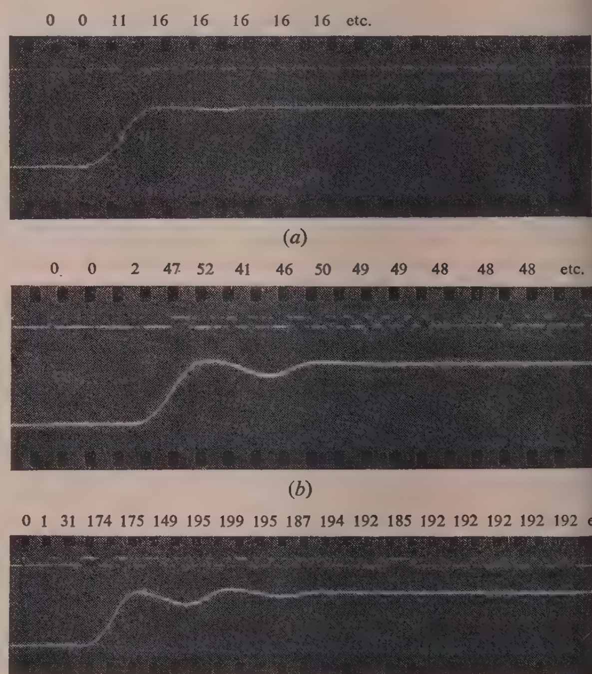
Table 3

NOISE RESPONSE IN A SIMPLE DIGITAL SERVO SYSTEM

Speed	Estimated input noise, N	Noise on error signal, N_e	$\sqrt{(3)}N$	No. of readings
deg/s				
7·75	0·75	1·45	1·30	211
14·6	1·3	2·8	2·25	112
22·7	2·1	4·6	3·63	148
30·8	2·9	6·33	5·02	118
34·8	3·3	5·56	5·72	141
40·9	4·0	7·26	6·94	119

(5.2.3) Step-Function Response.

The speed of rotation of the output shaft is a maximum at the mid-point of the transient, and the scale-reading error will be greatest at this point. The number is actually ahead of the true position, so that an apparent positive error will appear when there should be none. This positive error is responsible for a negative dip or "bounce" corresponding to the impulse response of the system to this error. If the speed is still high during this secondary transient, further oscillations will follow. Fig. 21



(c)

Fig. 21.—Response of simple digital servo to step functions.
(a) Step of 16 units.
(b) Step of 48 units.
(c) Step of 192 units.

illustrates the effect. The main trace shows the output voltage as obtained from the potentiometer attached to the output shaft. The secondary trace is the digit sequence constituting the output number. The synchronizing digit, which is always present as a positive pulse, marks the beginning of each number group and may be identified by the fact that it is opposite the brightening pulse marking the sampling instant on the main trace. The decimal equivalents of the binary numbers have been inserted for ease of interpretation, and the effect of the scale-reading errors is clearly observable.

(5.3) Digital Zero-Velocity-Lag Servo

he addition of the integral-of-error component needed to uce the velocity lag to zero makes the servo system much less ble, and increases the effects of quantizing and scale-reading ors.

.1) Stability in the Rest Condition.

system with integral-of-error control is unstable with the in feedback loop opened. The output wanders steadily in direction or the other, a condition which effectively ains within the limit of one element of code. The out- tends to oscillate in an erratic manner between numbers one ove and one below the correct zero position, at a frequency inly dependent upon the amount of integral control included. . 22 shows the type of oscillation experienced with the system up for $a = 0\cdot2$. The upper trace is the digital error as it is ilable from the subtractor. The output of the decoder is played in the lower trace and indicates that the jitter is fined to +1 and −1 unit of error.

Velocity lag and noise magnification may be measured together, ile the datum shafts are in continuous rotation. The former uld ideally be zero. Actually, the integrator in the A_2-loop is t perfect but has a decay time-constant, T, so that its transfer ction is $(1 + pT)^{-1}$. The velocity lag depends mainly upon T, d may be shown to be

$$\text{elocity lag} = \left[\frac{A_1(1+h)}{1+B_1+B_2} + \frac{A_2}{1-d}\right]^{-1} \quad \text{where } d = \varepsilon^{-T/\tau}$$

e results of a number of measurements are recorded in Table 4, d the agreement with the theory is reasonably good.

Table 4

VELOCITY LAG AND NOISE: COMPARISON OF MEASUREMENT WITH THEORY

Speed	Elements lag		Input noise, N	Error noise, N_e		No. of readings
	Meas.	Calc.		Meas.	Calc.	
$a = 0\cdot2$						
deg/s						
6·0	0·21	0·38	0·62	0·87	1·5	186
10·9	1·01	0·70	0·95	1·94	2·3	136
16·0	1·40	1·02	1·4	2·81	3·4	197
20·9	1·59	1·39	1·9	3·19	4·6	152
30·6	2·17	1·96	2·8	5·78	6·8	211
$a = 0\cdot4$						
6·0	0·65	0·69	0·62	0·95	1·2	162
10·9	1·39	1·25	0·95	1·71	1·8	128
16·0	1·93	1·84	1·4	2·69	2·7	191
20·9	2·35	2·40	1·9	2·66	3·7	143
30·6	3·33	3·50	2·8	5·37	5·4	100

The oscillation is violent but less than in Fig. 23(*b*) for which the step is 64 units. Much larger steps caused sustained oscillations. In Fig. 24, however, there is only a slight tendency to oscillate, and the response is very little worse than that of the analogue equivalent.

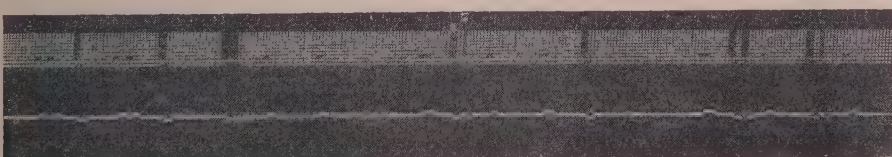


Fig. 22.—Stability in rest condition of digital servo having zero velocity lag.

Upper trace: Digital representation of error.
Lower trace: Analogue representation of error.

The r.m.s. noise on the error signal is again given by $= N[2(1 + M^2)]^{\frac{1}{2}}$, but since smoothing is included, the noise- wer gain is a function of a, namely

$$M^2 = \frac{1-a}{8(1+a)^5}(39 + 34a + 54a^2 + 34a^3 + 7a^4)$$

comparison of experimental results with theory is also included Table 4.

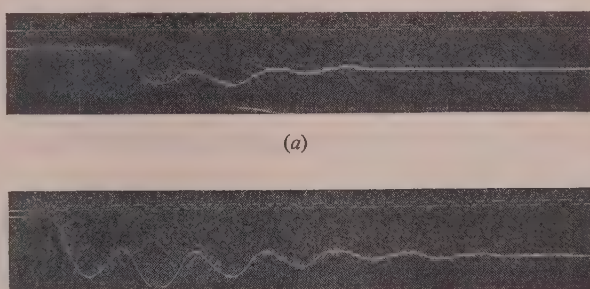
As was noted with the servo operating on analogue data, ter agreement with theory is obtainable with the larger ount of smoothing.

.3) Step-Function Response.

The most significant result which emerged during step-function ing of the zero-velocity-lag digital servo was that, under tain conditions, a large step could lead to sustained oscillations. e amplitude of step needed to produce this effect depended on the degree of smoothing. If $a = 0$ the system was so prone oscillation as to be virtually unusable. On the other hand, if $= 0\cdot4$ the smoothing was sufficient to damp out the effect of le-reading errors quite quickly.

The point is illustrated by Figs. 23 and 24, for which $a = 0\cdot2$, $a = 0\cdot4$ respectively. In Fig. 23(*a*) the size of step is 32 units.

The origin of the large or sustained oscillation when a is small lies in the fact that the first "bounce" caused by the first error in scale reading occurs at a time just following the peak of the overshoot. Consequently the velocity of return following the overshoot is increased and the "bounce" may be sufficient

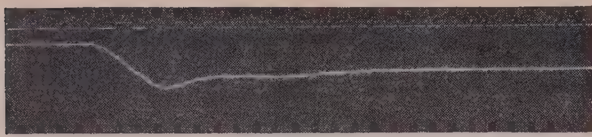


(*a*)

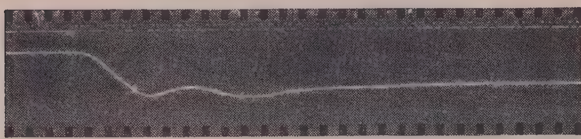
(*b*)

Fig. 23.—Response of digital servo having zero velocity lag to step functions ($a = 0\cdot2$).

(*a*) Step of 32 units.
(*b*) Step of 64 units.



(*a*)



(*b*)

Fig. 24.—Response of digital servo having zero velocity lag to step functions ($a = 0\cdot4$).

(*a*) Step of 32 units.
(*b*) Step of 128 units.

to cause a second large reading error. The process continues in a similar manner. When a is large, however, the first "bounce" occurs before the overshoot peak is reached and tends to cancel it out. The stability of the system seems in fact to be slightly improved by the presence of scale-reading errors. This would not be the case if the errors were not always in the leading direction. The general conclusion is that this type of servo would be satisfactory provided smoothing corresponding at least to $a = 0\cdot4$ were incorporated. Since lesser degrees of smoothing would in any case be accompanied by large overshoots and excessive noise magnification, it is not a serious limitation.

(6) CONCLUSIONS

Digital systems involve sampling and quantizing. To the first of these operations the principle of superposition still applies and a linear theory is possible. A method of synthesis has been suggested which takes into account the delay in the feedback loop likely to occur if relatively low-speed serial types of data-transducers and digital equipment are used. The method has been illustrated by two examples, in one of which summated error control is included to reduce the velocity lag to zero.

The experimental work undertaken has been directed first towards verifying the theory, and secondly towards studying the effects of digital encoding. The peculiarities have been deliberately exaggerated in order that they should not be masked by effects common to normal servo systems. Thus the data period is long compared with the time-constants of the system, the digit rate is low so that delays of one whole data-period are involved, the rate of scanning in the data transducer is slow so that large dynamic reading-errors are obtained and, finally, the amplitude quantizing (coding) is much coarser than it need be. In this way, evidence has been obtained which should be of value in designing according to established principles a practical system in which these defects are minimized as far as is consistent with other factors.

The peculiar effects due to quantizing and scale-reading errors have been studied, and it is possible to make estimates, based on the p.t.f. of the system, of the noise or distortion so introduced. Such estimates are reasonably well in agreement with the results of measurement, and it is apparent that digital systems with small amounts of smoothing are not likely to be very satisfactory in this respect. An important advantage results from the fact that scale-reading errors are always in the leading direction, for it should then be possible to adjust the smoothing so that these errors tend to increase the stability rather than to reduce it. On the other hand, they may, if the amount of smoothing is small, render unstable an otherwise stable system.

The general conclusion is that it is possible to construct digital servo systems which are capable of prediction to the extent likely to be required for making good the delay inherent in a digital data-transmission system. Since the stability diminishes and the sensitivity to noise and instrumental errors increases as the time of prediction is increased, it is essential that the delay in the transmission of each sample should be as small as possible. Furthermore, the sampling servo system is less efficient than a continuous system as a smoothing device, and as much smoothing as is practicable should be carried out at the sending end before the data are sampled.

(7) ACKNOWLEDGMENTS

Acknowledgment is made to the Controller of H.M. Stationery Office for permission to publish the paper. The work was carried out at the Signals Research and Development Establishment, Ministry of Supply, and the author wishes to thank the Chief Superintendent for permission to use it as material in a thesis for the degree of Doctor of Philosophy at London University.

(8) REFERENCES

(1) Barker, R. H.: "A Transducer for Digital Data-Transmission Systems" (see page 42).
(2) Barker, R. H.: "Group Synchronising of Binary Digital Systems: Communication Theory" (Butterworth, 1953), p. 273.
(3) Hazen, H. L.: "Theory of Servomechanisms," *Journal of the Franklin Institute*, 1934, **218**, p. 279.
(4) Gardner, M. F., and Barnes, J. L.: "Transients in Linear Systems" (John Wiley, 1942).
(5) MacColl, L. A.: "Fundamental Theory of Servomechanisms" (Van Nostrand, 1945), p. 88.
(6) Hurewicz, W.: "Filters and Servo Systems with Pulsed Data," M.I.T. Radiation Laboratory Series, **25** (McGraw-Hill, New York, 1947), p. 231.
(7) Lawden, D. F.: "A General Theory of Sampling Servo Systems," *Proceedings I.E.E.*, Monograph No. 4, July, 1951 (**98**, Part IV, p. 31).
(8) Barker, R. H.: "The Pulse Transfer Function and its Application to Sampling Servo Systems," *ibid.*, Monograph No. 43, July, 1952 (**99**, Part IV, p. 302).
(9) Ragazzini, J. R., and Zadeh, L. A.: "The Analysis of Sampled-Data Systems," *Transactions of the American I.E.E.*, Part II, p. 225, 1952.
(10) Linvill, W. K., and Salzer, J. M.: "Analysis of Control Systems involving Digital Computers," *Proceedings of the Institute of Radio Engineers*, **41,** 1953, p. 901.
(11) Salzer, J. M.: "Frequency Analysis of Digital Computers Operating in Real Time," *ibid.*, 1954, **42,** p. 457.
(12) Bergen, A. R., and Ragazzini, J. R.: "Sampled-Data Processing Techniques for Feedback Control Systems," *Transactions of the American I.E.E.* In course of publication.
(13) Lawden, D. F.: "Automatic and Manual Control" (Butterworth, 1952), p. 377.
(14) Whiteley, A. L.: "Theory of Servo Systems, with Particular Reference to Stabilization," *Journal I.E.E.*, 1946, **93**, Part II, p. 353.
(15) Williams, F. C., Robinson, A. A., and Kilburn, T.: "Universal High-Speed Digital Computers: Serial Computing Circuits," *Proceedings I.E.E.*, Paper No. 1164 M, April, 1952 (**99**, Part II, p. 107).

.3.012.8 : 621–52 : 628.8

Paper No. 1934 M
Jan. 1956

ELECTRICAL ANALOGUES FOR HEAT EXCHANGERS

By R. L. FORD, B.Sc., Ph.D., Graduate.

(*The paper was first received* 21*st February*, 1955, *and in revised form* 8*th August*, 1955.)

SUMMARY

The paper is concerned with the derivation of dynamically accurate ctrical analogues for heat exchangers with a view to their appli- ion in the solution of automatic-control problems relating to the er. The analogues are based on idealized equations describing the haviour of the thermal processes involved.

Two heat exchangers, both distributed-parameter systems, are con- ered, and two analogues are derived for each. One type of analogue composed almost entirely of passive components but the other ntains electronic amplifiers and employs a feedback technique.

The design and construction of a passive-network analogue is scribed, and its experimentally-determined frequency responses are en and compared with the theoretical responses of the heat exchanger. nally, an example is given of the experimental application of the alogue in an automatic-control loop.

LIST OF PRINCIPAL SYMBOLS

Thermal dimensions are given for most of the parameters d variables listed below, but in the text which follows, some these symbols are used to represent the analogous electrical antities. For example, depending upon the context, θ may interpreted as representing temperature or voltage. In the terests of simplicity no conversion factors have been employed such cases.

x = Distance along the heat-exchanger tube, cm.
l = Total immersed length of tube, cm.
$X = x/l$.
U = Mean circumference of tube, cm.
$, \alpha_2$ = Heat-transfer coefficients at inner and outer surfaces of tube, respectively, cal/cm² sec °C.
$\alpha = \alpha_1\alpha_2/(\alpha_1 + \alpha_2)$, overall heat-transfer coefficient for the tube, cal/cm² sec °C.
w_1 = Thermal capacitance of the tube fluid per unit length, cal/cm °C.
w_2 = Thermal capacitance of the tube per unit length, cal/cm °C.
w_3 = Thermal capacitance of the outer fluid per unit length of tube, cal/cm °C.
v_1 = Velocity of fluid flow in the tube, cm/sec.
v_2 = Effective velocity of outer fluid flow, cm/sec.
$\rho = v_1/v_2$.
$a_1 = U\alpha l/v_1w_1$.
$a_2 = U\alpha l/v_2w_3$.
t = Time, sec.
$\tau = tv_1/l$.
i = Heat flow, cal/sec.
θ = Temperature of fluid in tube; θ_i for inlet, θ_0 for outlet, °C.
T = Temperature of outer fluid; T_i for inlet, °C.
θ_T = Temperature of tube, °C.
f = Frequency, c/s.
ω = Angular frequency, rad/sec.
$\Omega = \omega l/v_1$.
s = Laplace transform operator with respect to time.
n = Number of sections in the electrical analogue.
$r = 1, 2, 3 \ldots n$.
θ_r = Output of the rth section of the electrical analogue, volts.
μ = Proportional control factor.
D = Derivative action time, sec.
I = Integral action time, sec.
ζ = Damping factor of a sinusoid whose amplitude decays exponentially.

(1) INTRODUCTION

Most heat exchangers are distributed-parameter systems and, as a result, their transfer functions are rather complex. This complexity is increased when the transfer function is combined with that of a process controller in order to investigate the dynamic behaviour of the system under closed-loop conditions. However, an electrical analogue of the heat exchanger used in conjunction with an analogue of the controller would yield the required results with a minimum expenditure of time and effort. The analogues described below were developed with this end in view.

It is not necessary to know the differential equations for a system in order to derive its analogue, but the ability to do so implies that the equations could be obtained if desired. Furthermore, for a linear system, if numerical values can be assigned to the components in an analogue, then a numerical solution of the equations could be obtained. It might be argued, therefore, that an automatic computer could be constructed to solve control problems relating to heat exchangers and accomplish this with great accuracy. This is probably true, but the computer would bear no resemblance to the heat exchanger and the only quantities which could be directly set into the machine would be the coefficients occurring in the differential equations. An analogue provides a physical picture of the system, and when the value of a component is adjusted we know immediately what this means in terms of the heat exchanger itself. This facility considerably simplifies an experimental approach towards improving the design of a system. Another advantage of the analogue is the ease with which most of the variable quantities can be monitored simultaneously. And having regard to accuracy, the quality of the data available for a heat exchanger is unlikely to be such as to justify the expense of an accurate computer.

Two types of electrical analogue will be described, each having their particular advantages depending on the application and the accuracy required. One of these is composed almost entirely of passive components but the other contains electronic amplifiers and employs a feedback technique. A passive-network analogue was built for one of the two heat exchangers which were considered and its experimentally-determined frequency responses are given and compared with the theoretical responses of the heat exchanger. Finally, an example is given of the experimental application of the analogue in determining the optimum controller settings required for the control of the outlet temperature of one of the fluids in the heat exchanger.

Written contributions on papers published without being read at meetings are ted for consideration with a view to publication.
Dr. Ford was formerly in the Electrical Engineering Dept., King's College, wcastle upon Tyne.

(2) CONSTANT-JACKET-TEMPERATURE HEAT EXCHANGER

Fig. 1(*a*) shows schematically the first heat exchanger to be considered. It consists of a tube, which carries a fluid flowing with velocity v_1, immersed in a constant-temperature jacket.

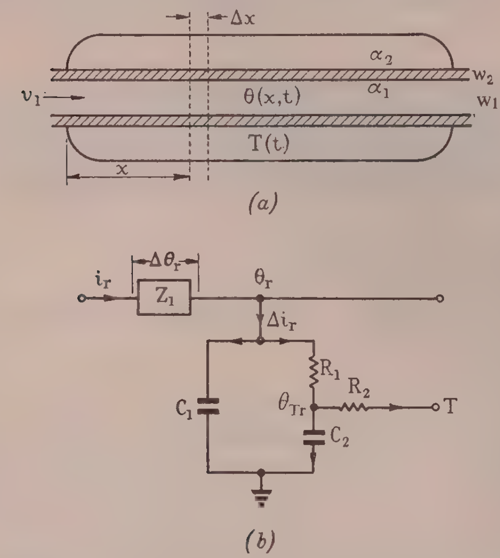


Fig. 1.—The constant-jacket-temperature heat exchanger.

(*a*) Schematic.
(*b*) Analogue of an element of length Δx.

The jacket temperature is constant only in the sense that it is not affected by the heat which is received from, or lost to, the tube and the inner fluid; it will be assumed that it can be made a function of time. This simple configuration is not particularly representative of a practical heat exchanger but it is useful as an introduction to more elaborate systems.

An initial study of the problem of simulating the process would be made unnecessarily complicated by trying to account for all the possible effects in the heat exchanger, and the following assumptions were therefore adopted:[1]

(*a*) Streamline flow of the fluid in the tube with the velocity and temperature uniform over any cross section.
(*b*) Heat transfer by conduction takes place in radial directions only. That is, the inner fluid and the tube have zero conductivity in the axial direction.
(*c*) The tube has infinite conductivity in radial directions.
(*d*) The parameters α and w are independent of temperature.

With these assumptions it is shown in Appendix 9.1.1 that the differential equation for the system is as follows:

$$\frac{\partial^2\theta}{\partial t^2} + v_1\frac{\partial^2\theta}{\partial t\partial x} + b_1\frac{\partial\theta}{\partial t} + v_1 b_2\frac{\partial\theta}{\partial x} + b_3\theta = b_3 T \quad . \quad . \quad (1)$$

where
$$\left.\begin{aligned} b_1 &= \frac{U}{w_1 w_2}[\alpha_1(w_1 + w_2) + \alpha_2 w_1] \\ b_2 &= \frac{U}{w_2}(\alpha_1 + \alpha_2) \\ b_3 &= \frac{U^2\alpha_1\alpha_2}{w_1 w_2} \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (2)$$

It is shown in Appendix 9.1.2 that the transfer functions for sinusoidal disturbances are given by

$$\frac{\theta_0}{\theta_i} = \varepsilon^{-\left[\frac{(b_3-\omega^2)+j\omega b_1}{b_2+j\omega}\right]\frac{l}{v_1}} \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

when the temperature of the outer fluid is taken as the referenc temperature, i.e. $T = 0$, and

$$\frac{\theta_0}{T} = \frac{b_3}{(b_3 - \omega^2) + j\omega b_1}\left\{1 - \varepsilon^{-\left[\frac{(b_3-\omega^2)+j\omega b_1}{b_2+j\omega}\frac{l}{v_1}\right]}\right\} \quad . \quad (4$$

when $\theta_i = 0$. In the steady state, these expressions may b reduced and combined to give

$$\begin{aligned}\theta_0 &= \theta_i\varepsilon^{-\frac{b_3 l}{b_2 v_1}} + T\left(1 - \varepsilon^{-\frac{b_3 l}{b_2 v_1}}\right) \\ &= (\theta_i - T)\varepsilon^{-a_1} + T \quad . \quad . \quad . \quad . \quad . \quad . \quad (5\end{aligned}$$

For comparison with the response of the analogue it is con venient to write eqn. (5) in the form

$$\theta_0 = (\theta_i - T)\left(\varepsilon^{\frac{a_1}{n}}\right)^{-n} + T \quad . \quad . \quad . \quad . \quad (5a$$

(2.1) First Passive-Network Analogue

The basis for an electrical analogue for a thermal system i provided by the obvious analogy between voltage and tem perature and between current and heat flow. This implies tha thermal resistance is analogous to electrical resistance and so on

Since the heat exchanger is a distributed-parameter systen while the majority of easily adjusted electrical networks ar lumped-constant systems, it is apparent at once that approximat methods must be employed in order to obtain a convenient forn of analogue. By dividing the heat exchanger into a number o equal elements of length Δx, as indicated in Fig. 1(*a*), an imagining the temperature changes along the tube to occur i discrete steps, a lumped-constant electrical analogue can be use for each element. By suitably connecting these electrical section an analogue for the complete system is obtained whose accurac depends on the number of elements into which the heat exchange is divided.

Fig. 1(*b*) shows the electrical analogue of the *r*th element c the heat exchanger. R_1 and R_2 represent the resistances to hea transfer at the inner and outer surfaces of the tube, respectivel so that the voltage across them is $(\theta_r - T)$. The voltage, θ_{Tr}, a the junction of R_1 and R_2 is a measure of the temperature of th tube, and the connection of resistors between these points o adjacent elements would make the network more closel analogous to a practical heat exchanger where axial heat flo must take place in the tube. However, this would considerabl increase the complexity of the analogue and it has not bee investigated.

C_1 and C_2 represent the lumped thermal capacitances ove the length Δx of the inner fluid and the tube, respectively. Th energies stored depend only on the temperatures of the flui and the tube in the element, so that the capacitors are connecte between earth and the points at potentials representing thes temperatures.

The method of connecting the adjacent elementary analogue must be such as to produce voltage differences between them i order to simulate the temperature gradient along the inner flui In Fig. 1(*b*) this requirement has been satisfied by placing a serie impedance Z_1 between the sections. This impedance has n physical equivalent in the heat exchanger so that its natur cannot be determined by direct analogy, but it is clear that i value will be a function of the velocity of fluid flow in the tub

Apart from Z_1, the values of the components shown in Fig. 1(*l* are determined as follows:

C = capacitance of a length Δx of the inner flow or the tub

Therefore
$$\left.\begin{aligned} C_1 &= w_1\Delta x \\ C_2 &= w_2\Delta x \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad ($$
and

$$\frac{1}{R} = \frac{\text{heat flow}}{\text{temperature difference}}$$

$$= \frac{U \times \Delta x \times \alpha \times \text{temperature difference}}{\text{temperature difference}}$$

erefore

$$\left.\begin{aligned} R_1 &= \frac{1}{U\alpha_1 \Delta x} \\ R_2 &= \frac{1}{U\alpha_2 \Delta x} \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

1

That the analogue is of the correct form thus far may be nonstrated by comparing its differential equations with those the heat exchanger. It is shown in Appendix 9.1.1 that n. (1) is based on the equations

$$-v_1 w_1 \frac{\partial \theta}{\partial x} = w_1 \frac{\partial \theta}{\partial t} + \alpha_1 U(\theta - \theta_T) \quad . \quad . \quad . \quad (8)$$

$$w_2 \frac{\partial \theta_T}{\partial t} = \alpha_1 U(\theta - \theta_T) - \alpha_2 U(\theta_T - T) \quad . \quad . \quad (9)$$

e analogue will be composed of a finite number of sections, we must ensure that the equations describing their behaviour rrespond as closely as possible to the finite-difference forms of ns. (8) and (9). In terms of finite differences with respect to x, ns. (8) and (9) may be written

$$-v_1 w_1(\theta_r - \theta_{r-1}) = w_1 \Delta x \frac{\partial \theta_r}{\partial t} + \alpha_1 U \Delta x(\theta_r - \theta_{Tr}) \, . \quad (10)$$

$$w_2 \Delta x \frac{\partial \theta_{Tr}}{\partial t} = \alpha_1 U \Delta x(\theta_r - \theta_{Tr}) - \alpha_2 U \Delta x(\theta_{Tr} - T) \quad (11)$$

sing the relationships expressed by eqns. (6) and (7), eqns. (10) d (11) become

$$v_1 w_1 \Delta \theta_r = C_1 \frac{\partial \theta_r}{\partial t} + \frac{(\theta_r - \theta_{Tr})}{R_1} \quad . \quad . \quad . \quad (10a)$$

$$C_2 \frac{\partial \theta_{Tr}}{\partial t} = \frac{(\theta_r - \theta_{Tr})}{R_1} - \frac{(\theta_{Tr} - T)}{R_2} \quad . \quad . \quad . \quad (11a)$$

eqn. (10*a*) we have written $-\Delta\theta_r$, the change in θ over the rth crement of length Δx, instead of $(\theta_r - \theta_{r-1})$. It is now parent that if we have

$$\Delta i_r = v_1 w_1 \Delta \theta_r \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

en eqns. (10*a*) and (11*a*) describe the behaviour of the shunt m of the circuit of Fig. 1(*b*).

1.1) Feeding and Terminating.

If any cross-section of the inner fluid moves a distance δx in ne δt, then, since the heat stored in an element of the fluid of ngth δx is $w_1 \delta x \theta$, the axial flow of heat due to the motion of the id is given by

$$i = \frac{w_1 \delta x \theta}{\delta t}$$

follows that

$$i = v_1 w_1 \theta \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (13)$$

Eqn. (13) shows that the rate of supply of heat to the heat changer via the tube fluid is $v_1 w_1 \theta_i$. Therefore, by direct alogy, the electrical network should be fed with a constant rrent $i_1 = v_1 w_1 \theta_i$. By hypothesis, the temperature of the id in the outer vessel is not affected by the heat which it ceives from, or loses to, the tube so that, in the analogue, the urce of the voltage T should have zero output impedance.

Eqn. (13) also shows that the rate of flow of heat leaving the heat exchanger via the tube fluid is $v_1 w_1 \theta_0$. In the analogue this output current must develop a voltage θ_0 across the terminating impedance. It follows that the network should be terminated in a resistance R_t of value $1/v_1 w_1$.

(2.1.2) Determination of the Series Impedance.

If the analogue is fed with a constant current $v_1 w_1 \theta_i$ the input impedance must be such that this current develops a voltage θ_i between the input terminal and earth. This indicates that the input impedance should be $1/v_1 w_1$ $(= R_t)$. In these circumstances and with $T = 0$ the analogue reduces to a four-terminal network of characteristic impedance R_t, as shown in Fig. 2(*a*).

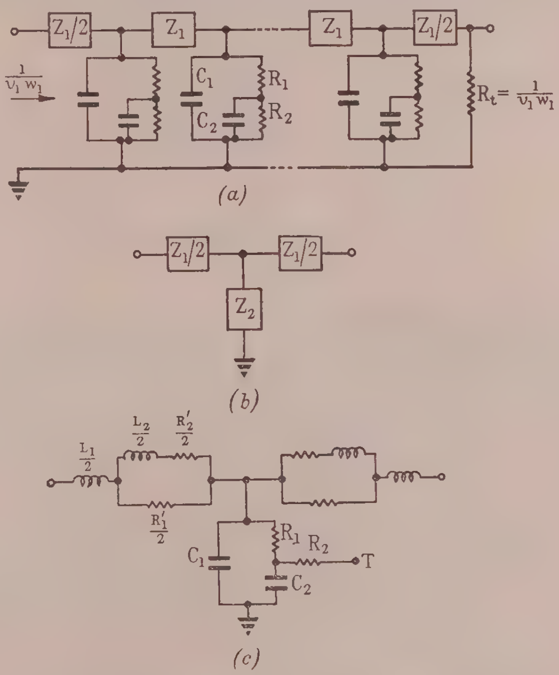


Fig. 2.—The analogue as a four-terminal ladder network.

(*a*) Complete analogue with $T = 0$.
(*b*) A T-section of the network with $T = 0$.
(*c*) Components of the T-section, $T \neq 0$.

Impedances $Z_1/2$ have been inserted at each end of the analogue, and the result is a four-terminal ladder network composed of T-sections of the form shown in Fig. 2(*b*). Each T-section is the analogue of one of the elements into which the heat exchanger has been divided.

It can now be seen that Z_1 should be such that the characteristic impedance of the network is R_t. The characteristic impedance Z_0 of the T-section of Fig. 2(*b*) is given by

$$Z_0 = \sqrt{\left[Z_1 Z_2\left(1 + \frac{Z_1}{4Z_2}\right)\right]} \quad . \quad . \quad . \quad . \quad (14)$$

Therefore, if $\sqrt{(Z_1 Z_2)} = 1/v_1 w_1$ and if $Z_1/4Z_2$ is much less than unity we have $Z_0 \simeq 1/v_1 w_1 = R_t$. Using the relationships given by eqns. (2), (6) and (7) it is easily shown that the shunt impedance Z_2 may be given by

$$Z_2 = \frac{1}{w_1}\left[\frac{b_2 + j\omega}{(b_3 - \omega^2) + j\omega b_1}\right]\frac{n}{l} \quad . \quad . \quad . \quad (15)$$

where $l/n = \Delta x$. It follows that we require

$$Z_1 = \frac{1}{v_1^2 w_1}\left[\frac{(b_3 - \omega^2) + j\omega b_1}{b_2 + j\omega}\right]\frac{l}{n} \quad . \quad . \quad . \quad (16)$$

Substituting the values for Z_2 and Z_1 from eqns. (15) and (16) in eqn. (14) we obtain

$$Z_0 = \frac{1}{v_1 w_1}\sqrt{\left\{1 + \frac{1}{4v_1^2}\left[\frac{(b_3 - \omega^2) + j\omega b_1}{b_2 + j\omega}\right]^2 \frac{l^2}{n^2}\right\}} \quad . \quad (17)$$

This shows that any desired degree of accuracy may be obtained by making n large enough.

Apart from the factor $1/v_1^2 w_1^2$, Z_1 is the inverse impedance of Z_2 and must take the form illustrated by Fig. 2(c), which shows the complete analogue of an element of length Δx of the heat exchanger. The values of the components of Z_1 are

$$\left.\begin{aligned} R_1' &= \frac{1}{v_1^2 w_1^2 R_1} \\ R_2' &= \frac{1}{v_1^2 w_1^2 R_2} \\ L_1 &= \frac{C_1}{v_1^2 w_1^2} \\ L_2 &= \frac{C_2}{v_1^2 w_1^2} \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad (18)$$

(2.1.3) Steady-State Response.

For constant direct currents and voltages the analogue as derived above reduces to the form shown in Fig. 3(a), where

$$\left.\begin{aligned} R' &= \frac{R_1' R_2'}{R_1' + R_2'} = \frac{1}{v_1^2 w_1^2 (R_1 + R_2)} \\ R &= R_1 + R_2 \end{aligned}\right\} \quad . \quad . \quad . \quad (19)$$

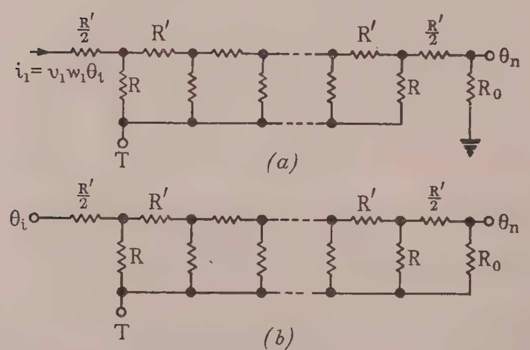


Fig. 3.—The analogue for constant, direct voltages and currents.
(a) Directly analogous methods of feeding and terminating the network.
(b) Correct methods.

Also, Z_0 as given by eqn. (17) reduces to a resistance R_0 given by

$$R_0 = \frac{1}{v_1 w_1}\sqrt{\left[1 + \left(\frac{a_1}{2n}\right)^2\right]} \quad . \quad . \quad . \quad . \quad (20)$$

It is convenient to terminate the analogue in R_0 rather than in R_t; in practice the difference should be very small.

Before proceeding further it can be seen from Fig. 3(a) that the unmodified analogue would give incorrect responses in two respects. If the network were fed as shown, the voltage appearing at the input terminal would be approximately $\theta_i + T$, whereas the desired voltage is θ_i. Also, we know that in the heat exchanger the output temperature can attain the value T, but this would not be correctly achieved by means of the circuit shown in Fig. 3(a). For example, if $\theta_i = T$ then, in the heat exchanger, $\theta_0 = T$, but this relationship would not obtain for the circuit of Fig. 3(a). These anomalies can be eliminated by feeding the network with a constant voltage θ_i and by returning R_0 to the source of T instead of to earth. The analogue is now as shown in Fig. 3(b) and for this circuit it is clear that when $\theta_i = T$ we have $\theta_n = T$.

It is shown in Appendix 9.2.1 that the response of the network of Fig. 3(b) to direct voltage inputs is given by

$$\theta_n = (\theta_i - T)\gamma_0^{-n} + T \quad . \quad . \quad . \quad . \quad (21)$$

where

$$\gamma_0 = \left\{1 + \frac{1}{2}\left(\frac{a_1}{n}\right)^2 + \frac{a_1}{n}\sqrt{\left[1 + \left(\frac{a_1}{2n}\right)^2\right]}\right\} \quad . \quad . \quad (22)$$

Comparing eqns. (21) and (5a) we see that the steady-state accuracy of the analogue depends only on the extent to which γ_0 is an adequate approximation to $\varepsilon^{a_1/n}$. Expanding these terms we have, if $a_1/n < 1$,

$$\varepsilon^{a_1/n} = 1 + \frac{a_1}{n} + \frac{1}{2!}\left(\frac{a_1}{n}\right)^2 + \frac{1}{3!}\left(\frac{a_1}{n}\right)^3 + \ldots$$

$$\gamma_0 = 1 + \frac{a_1}{n} + \frac{1}{2!}\left(\frac{a_1}{n}\right)^2 + \frac{3}{4}\left[\frac{1}{3!}\left(\frac{a_1}{n}\right)^3\right] + \ldots$$

The first three terms in the expansions are identical and the error is only 25% in the fourth, so that if a_1/n is small a high accuracy is possible; Table 1 illustrates the numerical values involved. When $a_1 = 2$ the attenuation in the heat exchanger is considerable, but a 5-section analogue would give adequate accuracy.

Table 1

STEADY-STATE ACCURACY OF THE PASSIVE-NETWORK ANALOGUE

a_1	ε^{-a_1}	$\left\{1+\frac{1}{2}\left(\frac{a_1}{n}\right)^2 + \frac{a_1}{n}\sqrt{\left[1+\left(\frac{a_1}{2n}\right)^2\right]}\right\}^{-n}$			% error			$\frac{R_1}{R_t}$
		$n = 2$	$n = 5$	$n = 10$	$n = 2$	$n = 5$	$n = 10$	$n = 5$
2	0·1353	0·1460	0·1371	0·1359	7·9	1·3	0·4	1·02
1	0·3679	0·3759	0·3685	0·3681	2·2	0·16	0·05	1·005

It can be shown that, when R_0 is returned to earth instead of to the source of T, the steady-state response of the network is given by

$$\theta_n = (\theta_i - T)\gamma_0^{-n} + T - \left[\frac{T}{2}(1 - \gamma_0^{-2n})\right] \quad . \quad . \quad (23)$$

It is evident that the spurious term gives rise to appreciable error in θ_n.

(2.1.4) Frequency Response.

In Appendix 9.2.2 it is shown that the response of the analogue to a sinusoidal disturbance in θ_i with $T = 0$ is given by

$$\theta_n = \theta_i \gamma^{-n} \quad . \quad . \quad . \quad . \quad . \quad . \quad (24)$$

where

$$\gamma = 1 + \frac{1}{2v_1^2}\left[\frac{(b_3 - \omega^2) + j\omega b_1}{b_2 + j\omega}\right]^2 \frac{l^2}{n^2} + \frac{1}{v_1}\left[\frac{(b_3 - \omega^2) + j\omega b_1}{b_2 + j\omega}\right]\frac{l}{n}\sqrt{\left\{1 + \frac{1}{4v_1^2}\left[\frac{(b_3 - \omega^2) + j\omega b_1}{b_2 + j\omega}\right]^2 \frac{l^2}{n^2}\right\}} \quad . \quad . \quad . \quad . \quad (25)$$

the derivation of this expression it was assumed that the ıalogue was terminated in Z_0. This is justified, as eqn. (17) ows, if n is large enough. Comparing eqns. (24) and (3) e see that the accuracy of the analogue depends on the greement between γ and $\varepsilon^{\frac{1}{v_1}\left[\frac{(b_3-\omega^2)+j\omega b_1}{b_2+j\omega}\right]\frac{l}{n}}$. Although the rst three terms in the expansions for these quantities are again entical, in this case the individual terms are functions of ω so at for a finite number of sections the analogue has a limited ındwidth.

Again assuming the network to be terminated in Z_0, it is shown Appendix 9.2.2 that the response to a sinusoidal disturbance T with $\theta_i = 0$ is given by

$$= \frac{b_3}{(b_3-\omega^2)+j\omega b_1}(1-\gamma^{-n}) - \frac{Z_1(Z-Z_2)(1+\gamma)}{4ZZ_0(1-\gamma)}(1-\gamma^{-2n}) \quad . \quad . \quad . \quad . \quad . \quad (26)$$

here
$$Z = \frac{1}{w_1}\left(\frac{b_2+j\omega}{b_3}\right)\frac{n}{l} \quad . \quad . \quad . \quad . \quad . \quad (27)$$

.eferring to eqn. (4) we see that the first term in eqn. (26) epresents the desired response and is subject to the same limits f accuracy as the response to a sinusoidal disturbance in θ_i. 'he second term arises from a mismatch in the termination of ıe analogue in so far as time-dependent values of T are conerned. The equation defining the behaviour of the last section f the analogue is not identical with the general equation which btains for the others, and this results in a spurious term in the esponse. When $\omega = 0$, $Z = Z_2$ and the spurious term vanishes.

.1.5) General.

Table 1 shows that the passive-network analogue can easily e made to give accurate results in the steady state. Unforınately, it is not possible to determine the accuracy for a ınusoidal disturbance in θ_i in such general terms. Even when ıitable values for the individual parameters are known or ssumed, eqn. (24) must be used with caution owing to the ssumption made in its derivation. It is clear, however, that as ong as ω is such that $Z_0 \simeq R_t$ the response should be reasonably ccurate.

Eqn. (26) indicates that the errors involved in the response to sinusoidal disturbance in T could be appreciable. However, was considered that the effect of the spurious term might be educed by terminating the analogue in a number of additional ections, i.e. by employing a total of $n + m$ sections but taking ıe output from the nth.

Assuming that the frequency-dependent errors are not proıibitive, there remain two inherent disadvantages in the analogue. he first is the necessity for inductors, components which are ot readily available in a standard range of values, and the econd is the fact that the value of each element of the series npedance is inversely proportional to v_1^2. It is therefore impraccable to make v_1 a continuous function of the output of a ontroller analogue for the purposes of automatic control. he analogue described in the next Section was developed in an ttempt to overcome these disadvantages.

(2.2) First Feedback Analogue

Fig. 4(a) shows schematically the basis for any lumpedonstant type of analogue for the heat exchanger when $T = 0$. he voltages θ_{r-1}, θ_r, etc., represent the temperatures of the ıbe fluid at intervals of length Δx along the tube. The problem to find a means of producing these voltages without recourse series impedances between sections.

Eqn. (12) provides the basis for an alternative method. If the correct voltage θ_r appears across the rth shunt impedance, the correct current Δi_r will flow through it. Assuming that this state of affairs exists, the voltage developed across a resistance $1/v_1w_1$ placed in series with Z_2, as shown in Fig. 4(b), will be $\Delta i_r/v_1w_1$ if $Z_2 \gg 1/v_1w_1$. From eqn. (12) we see that $\Delta i_r/v_1w_1 = \Delta\theta_r$, which is the required difference between the voltages across the $(r-1)$th and the rth impedance. Hence by a suitable arrangement of subtracting units, as shown in Fig. 4(c), these increments of voltage can be used to derive the voltages θ_{r-1}, θ_r, etc., which can then be applied to the appropriate impedances. In this way the original assumption that the correct values for θ_{r-1}, θ_r, etc., existed, is satisfied.

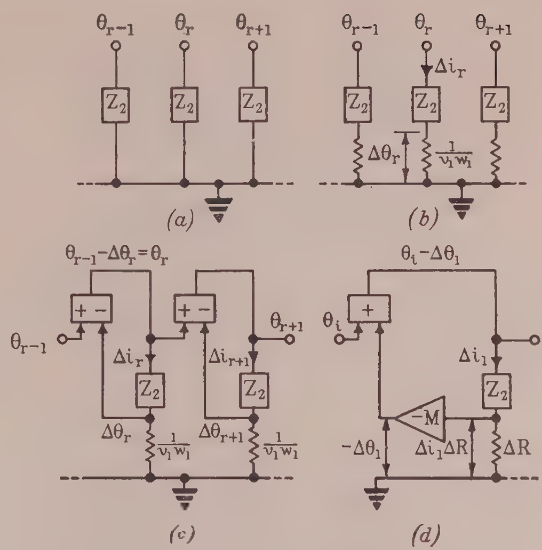


Fig. 4.—Development of the feedback analogue with $T = 0$.

$M = 1/v_1w_1\Delta R$.

Having developed the system thus far it can be seen that, since $\Delta\theta_r$ is the required voltage difference between the $(r-1)$th and rth sections, we cannot allow it to exist across the resistor in series with Z_2. However, errors from this source can be reduced to any desired proportions by replacing the series resistor by one of much smaller value (ΔR) and amplifying the voltage developed across it. This is illustrated by Fig. 4(d) which shows the first section of the feedback analogue. The voltage increment $\Delta i_1\Delta R$ is multiplied by the amplifier gain $-M$ to give $-\Delta\theta_1$. A phase-reversing amplifier is desirable so that negative feedback can be applied if required, and the subtracting unit is replaced by an adding unit.

(2.2.1) Introduction of *T*.

The method whereby the effect of the constant-temperature jacket can be introduced is illustrated by Fig. 5(a). The situation is complicated by the fact that while C_1 and C_2 are connected to earth, R_2 is returned to the source of the voltage T, so that Δi_1 is composed of two components $\Delta i_1'$ and $\Delta i_1''$ which cannot be passed through a common resistor. Neglecting, initially, the effect of introducing a resistance $1/v_1w_1$ in series with Z_2 we see from Fig. 5(a) that $\Delta\theta_1'$ $(= \Delta i_1'/v_1w_1)$ can be obtained directly, but T must be subtracted from the voltage appearing at A in order to obtain $\Delta\theta_1''$. Having done this, $\Delta\theta_1'$ and $\Delta\theta_1''$ must be added together to obtain $\Delta\theta_1$, which is then subtracted from θ_i to give θ_1.

A practical arrangement for effecting these operations is shown in Fig. 5(b). Some of the adding operations required

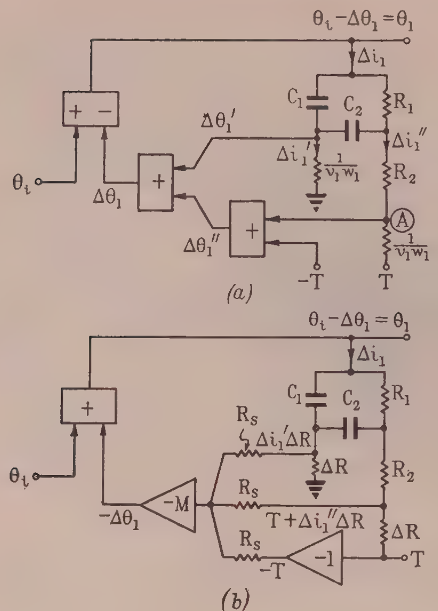


Fig. 5.—First section of the complete feedback analogue.

$M = 3/v_1 w_1 \Delta R$.

are carried out by means of the three resistances R_s, and the amplifier gain is increased to compensate for the attenuation of one-third which results. The adding unit for θ_i and $-\Delta\theta_1$ should have a low output impedance, must be non phase-reversing and should produce no attenuation. The feedback loop is inherently stable since the phase angle of Z_2 is always less than 180°.

Since the value of T is the same for all sections, the phase-reversing amplifier of unity gain which is shown in Fig. 5(*b*) can be made common to all sections. The output impedance of this amplifier must be much less than R_s, and that of the source of T must be very small in order to avoid the generation of spurious voltages.

(2.2.2) Response of the Feedback Analogue.

It is shown in Appendix 9.3 that the output voltage of the feedback analogue for constant, direct values of θ_i and T is given by

$$\theta_n = (\theta_i - T)\left(1 + \frac{a_1}{n}\right)^{-n} + T \quad . \quad . \quad . \quad (28)$$

In the derivation of this expression the effect of the series resistances ΔR was neglected.

Comparing eqns. (28) and (5*a*) we see that the steady-state accuracy of the feedback analogue depends on the agreement between the values for $(1 + a_1/n)$ and $\varepsilon^{a_1/n}$. This shows that the analogue is accurate only when a_1/n is so small that the exponential function can be adequately represented by the first two terms in the exponential series. It is evident, therefore, that in this respect the feedback analogue is considerably less accurate than a passive-network analogue having the same number of sections. Table 2 gives some numerical values for the errors in the feedback analogue, and we see that these compare unfavourably with those given in Table 1. However, although the values for n required to give comparable accuracies in the two cases are widely different it should be noted that, from a practical point of view, tolerable accuracy could be achieved with a feedback analogue if a_1 were not too large. For example, when

Table 2

STEADY-STATE ACCURACY OF THE FEEDBACK ANALOGUE

a_1	ε^{-a_1}	$\left(1+\frac{a_1}{n}\right)^{-n}$			% error		
		$n = 5$	$n = 10$	$n = 20$	$n = 5$	$n = 10$	$n = 20$
2	0·1353	0·1860	0·1614	0·1486	37·6	19·4	9·9
1	0·3679	0·4018	0·3855	0·3767	9·2	4·9	2·4
0·5	0·6066	0·6209	0·6138	0·6109	2·3	1·2	0·7

$a_1 = 0{\cdot}5$ Table 2 shows that a 5-section analogue would have an error of 2·3%, which is not excessive.

It is also shown in Appendix 9.3 that the responses of the analogue to sinusoidal disturbances in θ_i with $T = 0$ and in T with $\theta_i = 0$ are given by

$$\frac{\theta_n}{\theta_i} = \left\{1 + \frac{1}{v_1}\left[\frac{(b_3 - \omega^2) + j\omega b_1}{b_2 + j\omega}\right]\frac{l}{n}\right\}^{-n} \quad . \quad . \quad (29)$$

$$\frac{\theta_n}{T} = \frac{b_3}{(b_3 - \omega^2) + j\omega b_1}\left(1 - \left\{1 + \frac{1}{v_1}\left[\frac{(b_3 - \omega^2) + j\omega b_1}{b_2 + j\omega}\right]\frac{l}{n}\right\}^{-n}\right) \quad . \quad . \quad . \quad . \quad (30)$$

These expressions are similar in form to those defining the behaviour of the heat exchanger. Once again the theoretical accuracy of the analogue depends only upon the degree to which the exponential function can be satisfactorily approximated by the first two terms of the exponential series. Unlike the case of the passive-network analogue, eqn. (30) contains no spurious term due to mismatching.

(2.2.3) General.

It has been shown that an analogue which does not require inductors can be realized but only at the expense of increased complexity in other directions. More sections are required than would be necessary for a passive-network analogue of comparable accuracy. If ten sections be regarded as the practical limit, the numerical data given in Table 2 and the knowledge that errors increase with frequency indicate that the feedback analogue is of use only when a_1 is less than about 0·5. This assumes that an adequate bandwidth with frequency-dependent errors not exceeding, say, 10% could be achieved.

In the practical construction of a feedback analogue there are more possible sources of error than in the case of the passive-network analogue. For example, if ΔR were made small for minimum interference, M must be made large so that negative feedback could not be effectively applied to the amplifier unless it had a very high gain, and this implies the use of more than one stage of amplification. Also, however small ΔR is made, errors due to its presence will become apparent at frequencies where the reactance of C_1 is equally small.

The only parameters in the analogue which depend on v_1 are the amplifier gains. It would be quite practicable to vary the gains manually, and it should be possible to make them a function of a time-dependent voltage for the purpose of automatic control.

(3) ONE-FLUID-MIXED HEAT EXCHANGER

With the experience gained in deriving analogues for the constant-jacket-temperature heat exchanger we can now proceed to the investigation of a more practical system. The heat exchanger shown schematically in Fig. 6(*a*) might be regarded as the first complication of the constant-jacket-temperature heat exchanger. It consists of a tube AB immersed in a well-stirred

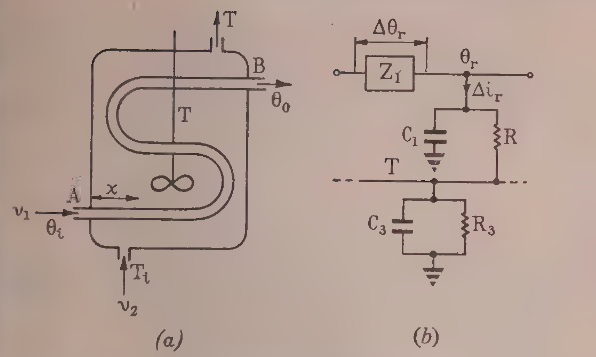


Fig. 6.—The one-fluid-mixed heat exchanger.

(a) Schematic.
(b) Analogue of an element of length Δx of the tube and the outer-flow analogue.

uid of finite thermal capacity, the flows being continuous hrough both the tube and the outer vessel.

Apart from the following additions, the assumptions made in rder to deal with this system were the same as those given in ection 2 for the constant-jacket-temperature heat exchanger:

(a) The thermal capacity of the tube itself may be neglected.
(b) No heat is lost from the surface of the outer vessel and the thermal capacity of the vessel itself may be neglected.
(c) The stirring is perfectly efficient so that the outer fluid is isothermal.

t was shown in Section 2.1 that the effect of the tube capacity an be simulated, and the reason for neglecting it here is simply o avoid unnecessary complication. The same applies to assumpion (b) as it will soon be apparent that these factors could easily e taken into account.

With the above assumptions it can be shown[1] that the ifferential equations for the system are as follows:

$$\frac{\partial\theta}{\partial t} + v_1\frac{\partial\theta}{\partial x} = \frac{U\alpha}{w_1}(T-\theta) \quad . \quad . \quad . \quad . \quad (31)$$

$$l\frac{dT}{dt} + v_2(T-T_i) = \frac{U\alpha}{w_3}\int_0^l(\theta - T)dx \quad . \quad . \quad (32)$$

hese may be written in dimensionless form as follows:

$$\frac{\partial\theta}{\partial\tau} + \frac{\partial\theta}{\partial X} = a_1(T-\theta) \quad . \quad . \quad . \quad . \quad (31a)$$

$$\rho\frac{dT}{d\tau} + (T-T_i) = a_2\left(\int_0^1\theta dX - T\right) \quad . \quad . \quad (32a)$$

t is shown in Appendix 9.4 that the transfer functions for he heat exchanger for sinusoidal disturbances in θ_i and T_i are iven by

$$\frac{\theta_0}{\theta_i} = \frac{[(1+j\rho\Omega)(a_1+j\Omega)^2 + ja_2\Omega(a_1+j\Omega)]\,\varepsilon^{-(a_1+j\Omega)} + a_1a_2 1 - [\varepsilon^{-(a_1+j\Omega)}]}{(1+j\rho\Omega)(a_1+j\Omega)^2 + ja_2\Omega(a_1+j\Omega) + a_1a_2[1-\varepsilon^{-(a_1+j\Omega)}]} \quad . \quad (33)$$

$$\frac{T}{\theta_i} = \frac{a_2(a_1+j\Omega)[1-\varepsilon^{-(a_1+j\Omega)}]}{(1+j\rho\Omega)(a_1+j\Omega)^2 + ja_2\Omega(a_1+j\Omega) + a_1a_2[1-\varepsilon^{-(a_1+j\Omega)}]} \quad . \quad . \quad . \quad (34)$$

hen $T_i = 0$, and

$$\frac{\theta_0}{T_i} = \frac{a_1(a_1+j\Omega)[1-\varepsilon^{-(a_1+j\Omega)}]}{(1+j\rho\Omega)(a_1+j\Omega)^2 + ja_2\Omega(a_1+j\Omega) + a_1a_2[1-\varepsilon^{-(a_1+j\Omega)}]} \quad . \quad . \quad . \quad (35)$$

$$\frac{T}{T} = \frac{(a_1+j\Omega)^2}{(1+j\rho\Omega)(a_1+j\Omega)^2 + ja_2\Omega(a_1+j\Omega) + a_1a_2[1-\varepsilon^{-(a_1+j\Omega)}]} \quad . \quad . \quad . \quad (36)$$

when $\theta_i = 0$. For constant values of θ_i and T_i the above equations may be reduced and combined to give

$$\theta_0 = \frac{[a_1\varepsilon^{-a_1} + a_2(1-\varepsilon^{-a_1})]\theta_i + a_1(1-\varepsilon^{-a_1})T_i}{a_1 + a_2(1-\varepsilon^{-a_1})} \quad . \quad (37)$$

$$T = \frac{a_2(1-\varepsilon^{-a_1})\theta_i + a_1T_i}{a_1 + a_2(1-\varepsilon^{-a_1})} \quad . \quad . \quad . \quad . \quad (38)$$

(3.1) Second Passive-Network Analogue

Fig. 6(b) shows the analogue of an element of length Δx of the heat-exchanger tube and inner flow, and also the equivalent electrical network for the fluid in the outer vessel. C_1 represents the lumped thermal capacitance of the tube fluid and R the total thermal resistance to heat transfer through the walls of the tube over the length Δx. As before, series impedances Z_1 are required in order to produce voltage differences between the elementary analogues.

Since the temperature of the outer fluid is not a function of x the resistances R are connected to a common point which is at a potential T. This potential is developed across C_3 and R_3 in parallel as shown in Fig. 6(b). C_3 represents the total thermal capacitance of the outer fluid, and the current through R_3 represents the flow of heat leaving the outer vessel via the outer fluid flow.

In terms of the heat-exchanger parameters, the components in the analogue, with the exception of Z_1, are as follows:

$$\left.\begin{aligned} C_1 &= w_1\Delta x \\ C_3 &= w_3 l \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad (39)$$

and
$$R = \frac{1}{U\alpha_1\Delta x} + \frac{1}{U\alpha_2\Delta x} = \frac{1}{U\alpha\Delta x} \quad . \quad . \quad . \quad (40)$$

The heat flow leaving the outer vessel is v_2w_3T so that R_3 is given by

$$R_3 = \frac{T}{v_2w_3T} = \frac{1}{v_2w_3} \quad . \quad . \quad . \quad . \quad . \quad (41)$$

v_2 and w_3 are, in fact, fictitious quantities owing to the physical arrangement of the heat exchanger but they are useful concepts.

The differential equations for the analogue are derived in Appendix 9.5, and it is shown there that in the limit, as Δx tends to zero, they reduce to eqns. (31) and (32) which define the behaviour of the heat exchanger.

(3.1.1) Completing the Analogue.

The heat flows entering the tube and the outer vessel at the inlets are $v_1w_1\theta_i$ and $v_2w_3T_i$, respectively, so that to be strictly analogous the passive network should be fed with constant currents as shown in Fig. 7(a). However, for the reason given in Section 2.1.3 for the constant-jacket-temperature heat-exchanger analogue, the tube analogue must be fed with a constant voltage θ_i as shown in Fig. 7(b). But since T is required

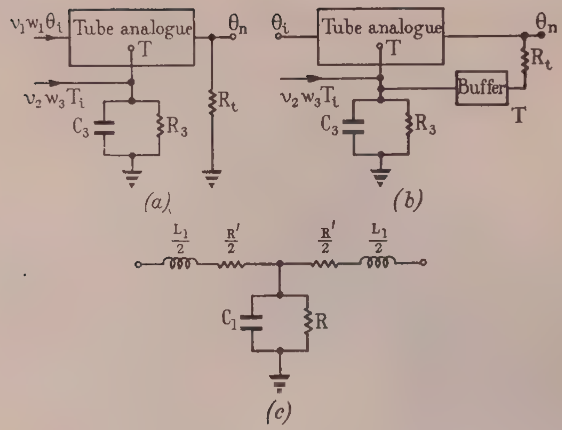


Fig. 7.—Feeding and terminating the analogue.

(a) Directly analogous methods.
(b) Correct methods.
(c) A T-section of the tube analogue.

to be a function of the current flowing into C_3 and R_3 from the tube analogue, the analogue of the outer fluid must be fed from a constant-current source.

The heat flow leaving the tube at the outlet is $v_1w_1\theta_0$, so that by direct analogy the tube analogue should be terminated in a resistance R_t of value $1/v_1w_1$, as shown in Fig. 7(*a*). As before, this would lead to errors in θ_n, and R_t must be returned to a point at a potential T, as shown in Fig. 7(*b*). An addition to the network in the form of a buffer stage is necessary in this case since the current in R_t must not be allowed to flow into C_3 and R_3.

Drawing on the experience gained in deriving the passive-network analogue for the constant-jacket-temperature heat exchanger, we can say immediately that the tube analogue is required to have a characteristic impedance Z_0 ($\simeq 1/v_1w_1$) as given by eqn. (14), Z_2 being the impedance of C_1 and R in parallel in this case. It follows that $Z_2 = R/(1 + j\omega RC_1)$ and substituting for C_1 and R from eqns. (39) and (40) we obtain

$$Z_2 = \frac{n}{v_1w_1}\left(\frac{1}{\dfrac{U\alpha l}{v_1w_1} + \dfrac{j\omega l}{v_1}}\right)$$

which may be written

$$Z_2 = \frac{n}{v_1w_1}\frac{1}{(a_1 + j\Omega)} \quad . \quad . \quad . \quad . \quad . \quad (42)$$

Hence, if

$$Z_1 = \frac{(a_1 + j\Omega)}{nv_1w_1} \quad . \quad . \quad . \quad . \quad . \quad . \quad (43)$$

the expression for Z_0 becomes

$$Z_0 = \frac{1}{v_1w_1}\sqrt{\left[1 + \left(\frac{a_1 + j\Omega}{2n}\right)^2\right]} \quad . \quad . \quad . \quad (44)$$

This shows that, if n is large enough, $Z_0 \simeq 1/v_1w_1$ as required, and when $\Omega = 0$ the expression reduces to eqn. (20).

Fig. 7(*c*) shows one of the T-sections of which the tube analogue must be composed in these circumstances, and the values for the components of Z_1 are as follows:

$$\left.\begin{aligned} R' &= \frac{1}{v_1^2w_1^2R} \\ L_1 &= \frac{C_1}{v_1^2w_1^2} \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (45)$$

(3.1.2) **Response of the Analogue.**

It is shown in Appendix 9.6 that the steady-state response is given by the following equations:

$$\theta_n = \frac{[a_1(v_1w_1R_0)\lambda_0^{-n} + a_2(1 - \lambda_0^{-n})]\theta_i + a_1(v_1w_1R_0)(1 - \lambda_0^{-n})T_i}{a_1(v_1w_1R_0) + a_2(1 - \lambda_0^{-n})} \quad . \quad . \quad . \quad . \quad (46)$$

$$T = \frac{a_2(1 - \lambda_0^{-n})\theta_i + a_1(v_1w_1R_0)T_i}{a_1(v_1w_1R_0) + a_2(1 - \lambda_0^{-n})} \quad . \quad . \quad . \quad (47)$$

The expression for λ_0 is identical to that for γ_0, which is given by eqn. (22), but it is convenient to use a different symbol for the one-fluid-mixed heat-exchanger analogue.

Comparing eqns. (46) and (47) with eqns. (37) and (38), respectively, we see that apart from the factor $v_1w_1R_0$ they are of the same general form. We already know that eqn. (22) can provide a good approximation to the exponential function. Therefore, since $v_1w_1R_0 = \sqrt{[1 + (a_1/2n)^2]}$, it follows that any desired accuracy can be obtained in the steady state by making n large enough.

In the derivation of the frequency response of the analogue, which is given in Appendix 9.6, it was assumed that the network was perfectly matched at its termination. Without this assumption, which holds exactly only in the steady state, the analysis is rather complicated. The transfer functions thus obtained are as follows:

$$\frac{\theta_n}{\theta_i} = \frac{(v_1w_1Z_0)[(1 + jp\Omega)(a_1 + j\Omega)^2 + ja_2\Omega(a_1 + j\Omega)]\lambda^{-n} + a_1a_2(1 - \lambda^{-n})}{(v_1w_1Z_0)[(1 + jp\Omega)(a_1 + j\Omega)^2 + ja_2\Omega(a_1 + j\Omega] + a_1a_2(1 - \lambda^{-n})} \quad . \quad . \quad . \quad . \quad (48)$$

$$\frac{T}{\theta_i} = \frac{a_2(a_1 + j\Omega)(1 - \lambda^{-n})}{(v_1w_1Z_0)[(1 + jp\Omega)(a_1 + j\Omega)^2 + ja_2\Omega(a_1 + j\Omega)] + a_1a_2(1 - \lambda^{-n})} \quad . \quad . \quad . \quad . \quad (49)$$

when $T_i = 0$, and

$$\frac{\theta_n}{T_i} = \frac{(v_1w_1Z_0)a_1(a_1 + j\Omega)(1 - \lambda^{-n})}{(v_1w_1Z_0)[(1 + jp\Omega)(a_1 + j\Omega)^2 + ja_2\Omega(a_1 + j\Omega)] + a_1a_2(1 - \lambda^{-n})} \quad . \quad . \quad . \quad . \quad (50)$$

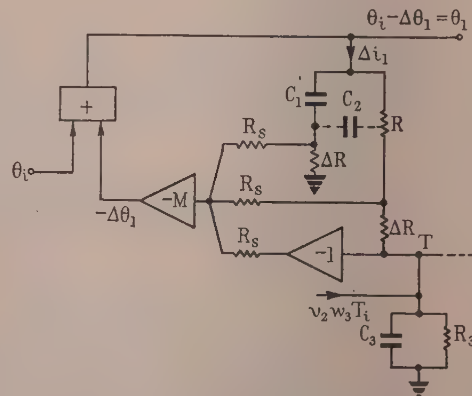


Fig. 8.—First section of a feedback analogue for the one-fluid-mixed heat exchanger.

$M = 3/v_1w_1\Delta R$.

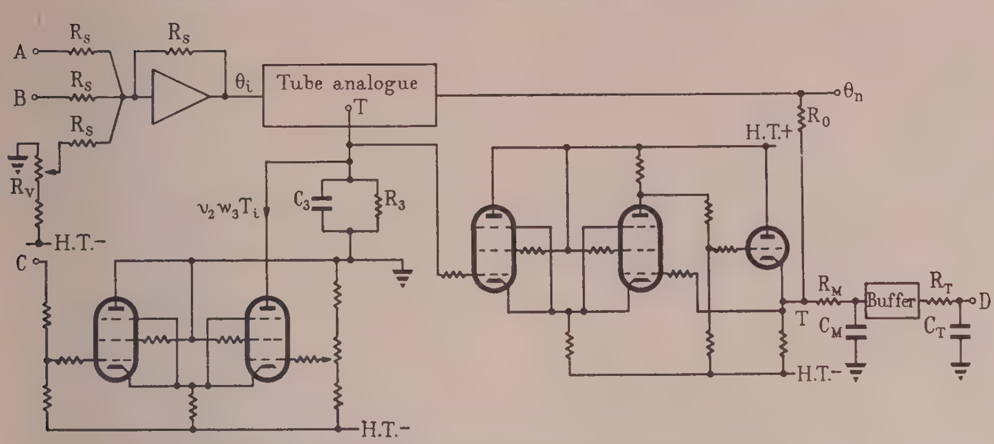


Fig. 9.—The passive-network analogue and its associated circuits.

$$= \frac{(v_1 w_1 Z_0)(a_1 + j\Omega)^2}{w_1 Z_0)[(1 + jp\Omega)(a_1 + j\Omega)^2 + ja_2\Omega(a_1 + j\Omega)] + a_1 a_2(1 - \lambda^{-n})} \quad . \quad . \quad . \quad . \quad (51)$$

en $\theta_i = 0$, and where

$$= \left\{1 + \frac{1}{2}\left(\frac{a_1 + j\Omega}{n}\right)^2 + \left(\frac{a_1 + j\Omega}{n}\right)\sqrt{\left[1 + \left(\frac{a_1 + j\Omega}{2n}\right)^2\right]}\right\} \quad (52)$$

is given by eqn. (44). These expressions indicate that the curacy of the frequency response depends only on the magude of $(a_1 + j\Omega)/n$, but we must bear in mind the fact that ey were derived on the assumption of a perfect termination.

(3.2) Second Feedback Analogue

An analogue of the feedback type, as described in Section 2.2, also possible for the one-fluid-mixed heat exchanger, but it is t proposed to discuss it in any great detail. The first section such an analogue is shown in Fig. 8. C_3 and R_3 and the 1 unit are common to all sections and must be fed from a nstant-current source as in the case of the passive-network alogue. Unlike the latter, however, it is easy to allow for e thermal capacity of the tube simply by introducing an ditional capacitance C_2 as indicated in Fig. 8. Similarly, the ect of tube conductivity in the axial direction could be simued by connecting resistors between the appropriate points on jacent sections, although a separate -1 unit would then be quired for each section.

Neglecting the effect of the series resistances ΔR, the responses the feedback analogue to a sinusoidal disturbance in θ_i are own in Appendix 9.7 to be given by the following equations:

$$= \frac{[(1 + jp\Omega)(a_1 + j\Omega)^2 + ja_2\Omega(a_1 + j\Omega)]\lambda_1^{-n} + a_1 a_2(1 - \lambda_1^{-n})}{(1 + jp\Omega)(a_1 + j\Omega)^2 + ja_2\Omega(a_1 + j\Omega) + a_1 a_2(1 - \lambda_1^{-n})} \quad . \quad . \quad . \quad . \quad (53)$$

$$= \frac{a_2(a_1 + j\Omega)(1 - \lambda_1^{-n})}{(1 + jp\Omega)(a_1 + j\Omega)^2 + ja_2\Omega(a_1 + j\Omega) + a_1 a_2(1 - \lambda_1^{-n})} \quad . \quad . \quad . \quad . \quad (54)$$

en $T_i = 0$, and where

$$\lambda_1 = \left[1 + \left(\frac{a_1 + j\Omega}{n}\right)\right] \quad . \quad . \quad . \quad . \quad (55)$$

e responses to a sinusoidal disturbance in T_i have not been termined. The only source of error in the above equations

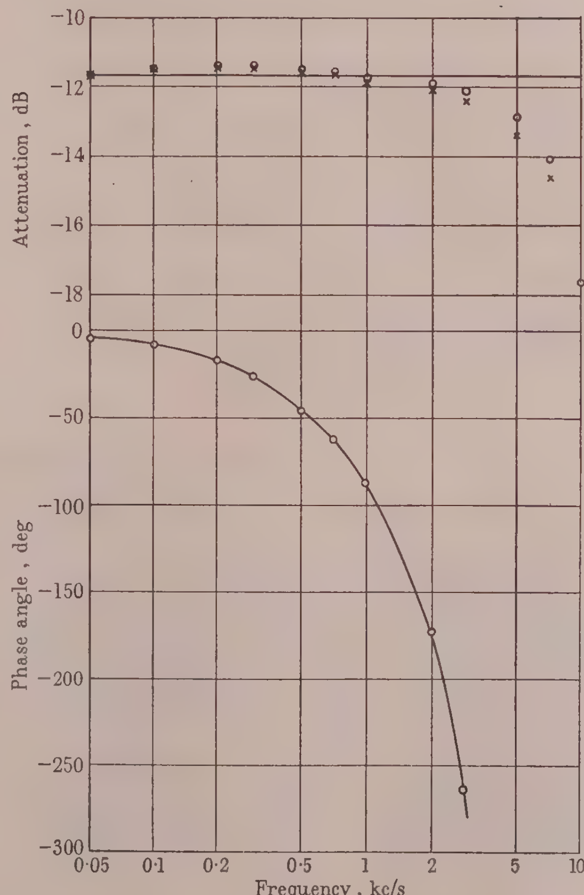


Fig. 10.—Frequency response of the tube analogue.

——— Heat-exchanger tube response.
○ ○ Analogue response; R_0 termination.
× × Analogue response; five-section unit termination.

lies in the approximation to the exponential term. As before, the analogue is considerably less accurate in this respect than the passive-network analogue but there is no matching problem.

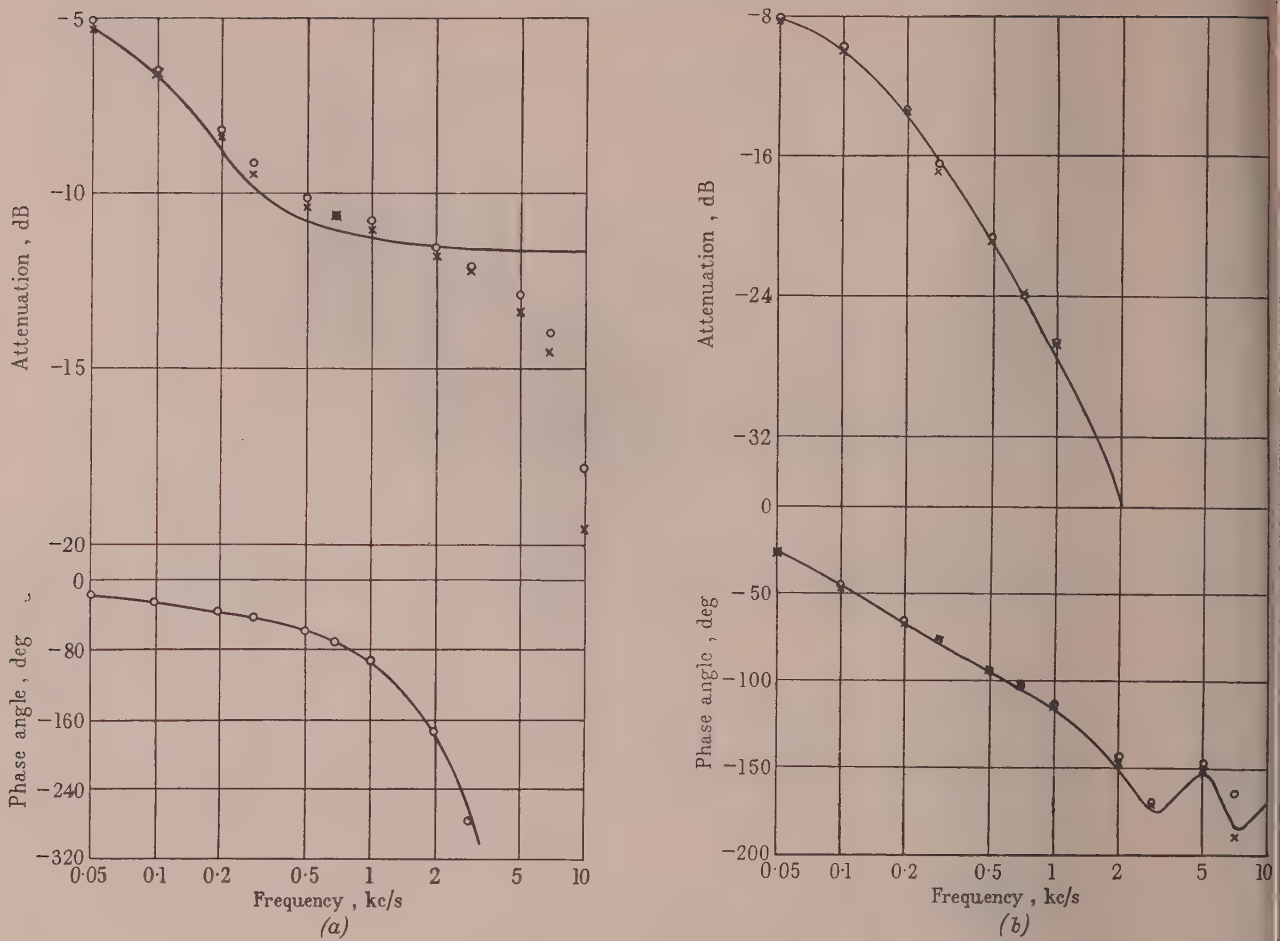


Fig. 11.—Frequency responses of the complete analogue.

(a) θ_n/θ_i. (b) T/θ_i. (c) θ_n/T_i. (d) T/T_i.

——— Heat-exchanger response.
○ ○ Analogue response; R_0 termination.
× × Analogue response; five-section unit termination.

(4) THE DESIGN AND TESTING OF AN ANALOGUE

Having developed the theoretical aspect of analogues for heat exchangers thus far, it seemed that an experimental investigation of the problem would be profitable. The one-fluid-mixed heat exchanger was decided upon, since it is the most realistic of the two heat exchangers which have been considered, and it only remained to choose between the two types of analogue. Owing to its greater accuracy in so far as the approximation to the exponential term is concerned, and since the effect of mismatching had not been determined, attention was concentrated on the passive-network analogue.

The design of the analogue was based on the following considerations:

(a) Typical parameters for the tube and inner fluid of a heat exchanger.
(b) The frequency range over which it was estimated that the analogue would be reasonably accurate.
(c) Available components.

No typical figures were available for the outer fluid but it was considered that, since the outer vessel must accommodate the tube and an efficient stirring mechanism, the thermal capacity of the outer fluid might well be ten times greater than that of the tube fluid. Its effective velocity was taken to be $v_1/10$. These assumptions resulted in $a_1 = a_2$, which was convenient for the purposes of calculation and at the same time quite practicable.

From the tube parameters it was estimated that the errors in the approximation to the exponential term for an analogue having ten sections would be less than 10% in amplitude and 10° in phase angle up to a frequency defined by $\omega = 0 \cdot 2$ rad/sec. The phase lag produced by the exponential term is about 400 at this frequency, and so it appeared that the analogue would be suitable for automatic-control experiments where phase lag of the order of 180° are of most interest. With this end in view a time-scale factor of 10^5 was chosen so that the resulting estimated bandwidth of the analogue was about 3 kc/s. This was then conveniently within the accurate range of an electrical analogue for a three-term controller which had already been constructed and which is described in a previous paper.[2] The fast time scale was also advantageous in that the transient behaviour of the system could be examined by means of an oscilloscope.

Having fixed the time scale, the scale factors for temperature and heat flow were chosen so that available inductors and capacitors could be utilized. In this process the heat-exchanger tube parameters were slightly modified to convenient values

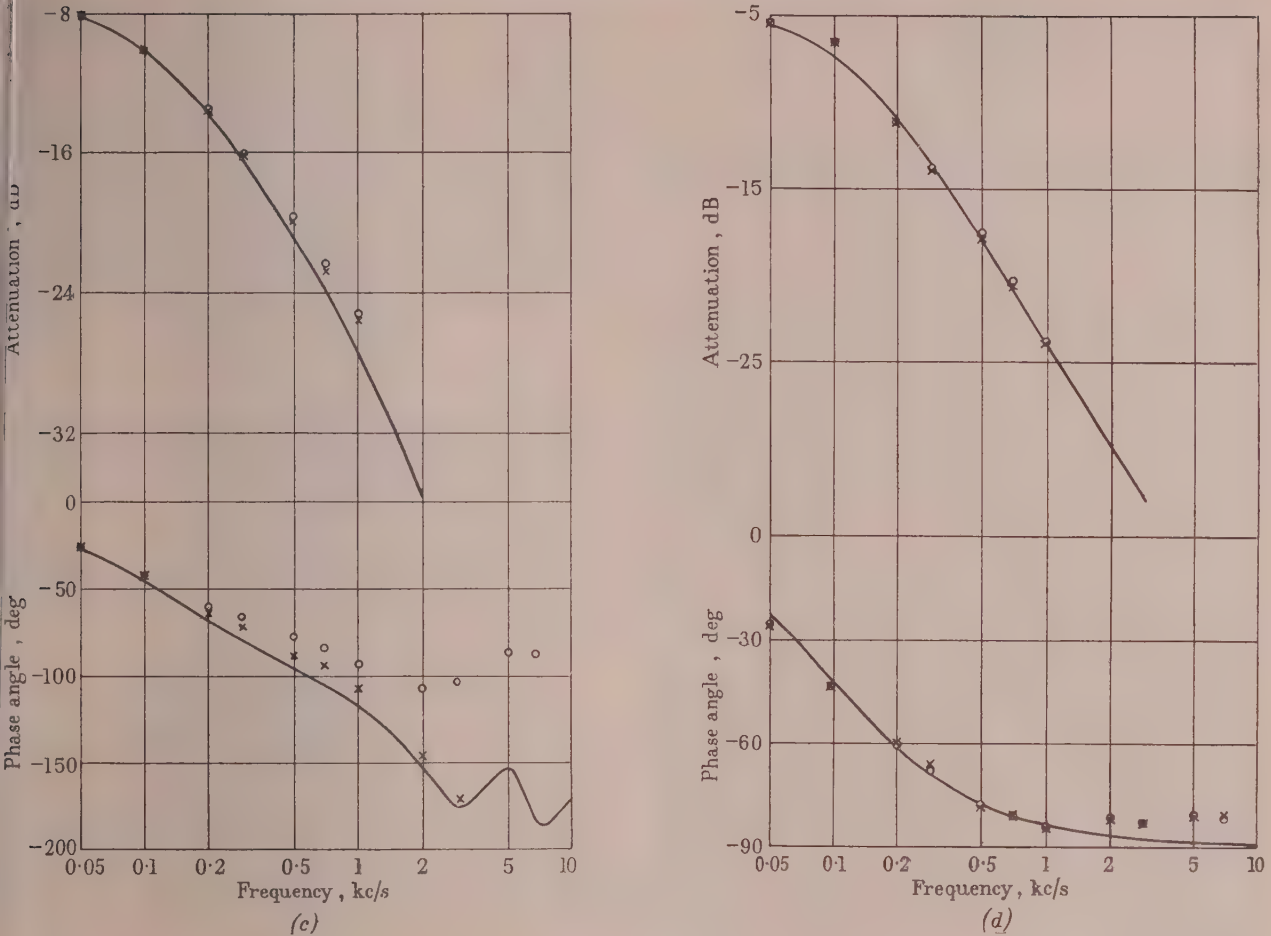


Fig. 11—*(continued)*

The resulting values of the components in the analogue were as ollows:

$$R = 3\cdot 65\,\text{k}\Omega \qquad C_1 = 0\cdot 05\,\mu\text{F}$$
$$R' = 65\cdot 8\,\Omega \qquad C_3 = 5\,\mu\text{F}$$
$$R_0 \simeq R_3 = 490\,\Omega \qquad L_1 = 12\,\text{mH}$$

The heat-exchanger parameters and the derivation of the scale actors are given in Appendix 9.8.

The series resistances R', being of low value, were wound from esistance wire, allowance being made for the d.c. resistance 3 ohms) of the inductors. The latter were high-grade comonents having a tolerance of $\pm 1\%$. Paper capacitors and omposition resistors were used for C_1 and R, respectively, and ll were selected to be within $\pm 1\%$ of their required values.

The tube analogue was composed of ten sections, and an dditional five-section unit was constructed in order to investigate he effects of using it as a termination as suggested in Section 2.1.5.

(4.1) Driving Circuits

Three units were required in order to supply the passiveetwork analogue with suitable voltages and currents, and the omplete system is shown in Fig. 9. The buffer stage, to prevent he current in R_0 from flowing into C_3 and R_3, was required to ave a gain of unity and, since $R_0 = 490$ ohms, a very low output impedance. These requirements were achieved, as shown in Fig. 9, by comparing the voltage T with the output of a cathode-follower and applying the amplified difference to the grid of the latter. In this way the cathode-follower output voltage is driven to be very nearly equal to T and the output impedance is made very small (2 or 3 ohms) by virtue of the feedback. It might be noted in passing that the input impedance of this circuit is very high since the grid-to-cathode voltage of the input valve is much less than the input voltage.

A further consequence of the relatively low impedance of the tube analogue was that θ_i had to be supplied from a low-impedance source. A high-gain d.c. amplifier, fed back to have a gain of -1 was employed for this purpose. Three input channels were provided; the steady-state value of θ_i was adjusted by means of the potentiometer R_v, sinusoidal or other disturbances were fed in at A while the controller-analogue output could be fed in at B if required.

A constant-current supply for the outer-fluid analogue was obtained by connecting C_3 and R_3 into one of the anode circuits of a cathode-coupled amplifier using pentode valves. One control grid was then available for manual adjustment of the constant current while a disturbance in T_i could be introduced as a voltage via the remaining grid, i.e. at C in Fig. 9.

The purpose of the networks composed of R_M, C_M and R_T, C_T, which are shown in Fig. 9, is discussed in Section 5.2.

(4.2) Experimental Results

From the data given in Appendix 9.8 we have

$$a_1 = a_2 = 1\cdot 34$$
$$\Omega = 24\cdot 5\omega$$
$$\rho = 10$$

Using the above value for a_1 and a_2 the evaluation of eqns. (37) and (38), which describe the steady-state response of the heat exchanger, gives

$$\left.\begin{aligned} \theta_0 &= 0\cdot 575\theta_i + 0\cdot 425T_i \\ T &= 0\cdot 425\theta_i + 0\cdot 575T_i \end{aligned}\right\} \quad . \quad . \quad . \quad (56)$$

It was found that the response of the analogue to constant, direct values of θ_i and T_i could be accurately predicted by means of eqns. (56).

The frequency responses of the heat exchanger were calculated from eqns. (33)–(36) and the curves obtained are shown in Figs. 10 and 11. Since $a_1 = a_2$, the transfer functions T/θ_i and θ_0/T_i are identical. The frequency scale has been multiplied by 10^5 to correspond with the analogue scale, and the experimentally determined responses of the analogue are indicated by the plotted points.

Fig. 10 shows the response of the tube analogue alone and therefore illustrates the accuracy of the approximation to the exponential term. When terminated in R_0 the attenuation characteristic was accurate to within 1 dB up to a frequency of $4\cdot 5$ kc/s, the errors in the phase characteristic being very small over this range. A slight hump in the attenuation characteristic in the region of 300 c/s was reduced in amplitude by terminating the analogue in the additional five-section unit, but this effected increased attenuation errors at the higher frequencies. The effect on the phase characteristic of this method of termination was negligible. It might be noted here that the phase lag at $3\cdot 2$ kc/s ($\omega = 0\cdot 2 \times 10^5$) is only 280° whereas a figure of 400° was quoted in connection with the estimation of the accuracy of the analogue. This apparent anomaly is due to the modification of the heat-exchanger tube parameters after the estimation had been made.

Figs. 11(*a*) and 11(*b*) show the responses of the complete analogue to a sinusoidal disturbance in θ_i. In the frequency band $0\cdot 2$–1 kc/s (corresponding to the band containing the hump in the tube-analogue attenuation characteristic), appreciable errors occurred in $|\theta_n/\theta_i|$. These were considerably reduced by terminating in the five-section unit but once again the higher-frequency attenuation was increased. The accuracy of the T/θ_i response was good with the analogue terminated in R_0, and some improvement in the phase characteristic at high frequencies was effected by means of the five-section termination.

Fig. 11(*c*) shows that large errors occurred in the response of θ_n to a sinusoidal disturbance in T_i, especially in the phase characteristic. These were reduced, but not to any particularly satisfactory extent, by terminating the tube analogue in the five-section unit. The response of T, shown in Fig. 11(*d*), was more accurate but this was probably due to the fact that it is mainly dependent on C_3 and R_3 and not to any marked extent, except at low frequencies, on the tube analogue.

In general, the results obtained indicated that some of the analogue transfer-functions were sufficiently accurate for automatic control purposes and this is considered in the next Section.

(5) APPLICATION OF AUTOMATIC CONTROL

In applying automatic control to the heat-exchange process either θ_0 or T may be the controlled condition, and these quantities may be controlled by regulating θ_i, T_i, v_1 or v_2. In the case of the passive-network analogue it was only practicable to regulate θ_i or T_i, although with a slower time scale it would have been possible to regulate v_2 by using an instrument servo-mechanism to vary R_3. Being limited in this way, the choice lay between the θ_0/T_i and the T/θ_i transfer functions. These were identical so far as the heat exchanger was concerned, but Figs. 11(*b*) and 11(*c*) showed that it was desirable to control T by regulating θ_i owing to the greater accuracy of the analogue in this case. Eqns. (49) and (50) show that, apart from the termination problem, the accuracy of the θ_n/T_i response would be expected to be less than that of the T/θ_i response owing to the presence of the factor $v_1 w_1 Z_0$ in the numerator of eqn. (50).

The equation describing the behaviour of a three-term process controller may, for the present purpose, be written

$$\Theta = -\mu\left[(T - T_0) + \frac{1}{I}\int (T - T_0)dt + D\frac{d}{dt}(T - T_0)\right] \quad . \quad (57)$$

where T_0 is the desired value of T. Θ has the dimensions of temperature and represents the difference between T_0 and the value which T would eventually attain if at any instant the control loop were opened—the regulating-valve* setting thereafter remaining constant—and the process allowed to reach equilibrium. Defining the controller behaviour in this way means that μ, which is in fact the steady-state open-loop gain, is a function of the attenuation in the heat exchanger. Eqns. (56) show that in the steady state and with $T_i = 0$ we have $T/\theta_i = 0\cdot 425$. It follows that

$$\mu = \text{Steady-state controller gain}\dagger \times 0\cdot 425 \quad . \quad . \quad (58)$$

For good control μ should be as large as possible, consistent with the requirement that the subsidence ratio of any oscillatory component in the closed-loop transient response to a step-function disturbance in θ_i or T_i should not be less than $\varepsilon : 1$. And after such a disturbance the value of the controlled quantity should return to the region of the desired value as rapidly as possible without appreciably undershooting the control point. These criteria for optimum control, which are arbitrary, are more fully discussed elsewhere.[3,4,2]

(5.1) Proportional Control

For proportional control $D = 0$ and I is infinite so that it is only necessary to determine a suitable value for μ. This would have been an extremely simple matter using the heat exchanger and controller analogues but in order to check the performance of the system the required value for μ was determined by calculation.

Owing to the large value for ρ, the term $(1 + j\rho\Omega)(a_1 + j\Omega)^2$ in the denominators of the heat-exchanger transfer functions is much larger than the others when $f > 0\cdot 01$ c/s.‡ In these circumstances a good approximation for the transfer function of eqn. (34) is as follows:

$$\frac{T}{\theta_i} \simeq \frac{a_2[1 - \varepsilon^{-(a_1 + j\Omega)}]}{(1 + j\rho\Omega)(a_1 + j\Omega)} \quad . \quad . \quad . \quad . \quad (59)$$

Fig. 11(*b*) shows that the phase lag at $0\cdot 01$ c/s is about 120° so that the lowest frequency of oscillation is certain to lie in the band where this approximate transfer function is valid.

In terms of the Laplace transform, eqn. (59) may be written

$$G_1(s) \simeq \frac{a_2[1 - \varepsilon^{-(a_1 + sL)}]}{(1 + s\rho L)(a_1 + sL)} \quad . \quad . \quad . \quad . \quad (60)$$

* In the heat-exchanger control loop the controller output might be employed to position a mixing valve in order to regulate θ_i. Assuming a linear characteristic for this valve, it need not be simulated in the analogue of the control loop.

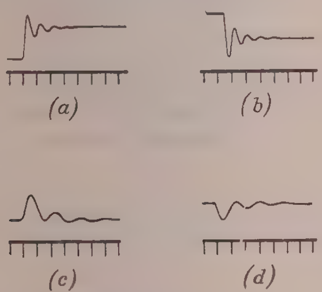
† This includes the gain of the regulating valve.

‡ This and all subsequently mentioned frequencies and periods are given in terms of the heat-exchanger time scale.

here $$G_1(s) = \mathscr{L}\left[\frac{T}{\theta_i}\right]$$

$$L = l/v_1$$

y plotting the attenuation and phase characteristics of $G_1(s)$ or $s = [-\zeta + j\surd(1-\zeta^2)]\omega$ we obtain the response of the eat exchanger to an exponentially-decaying sinusoid of subidence ratio $[\varepsilon^{2\pi\zeta/\surd(1-\zeta^2)}]:1$. If we take $\zeta = 0{\cdot}157$ the subidence ratio is $\varepsilon:1$, and from the phase characteristic we an determine the frequency at which the phase lag first ecomes 180° in these circumstances. The reciprocal of the nodulus of $G_1(s)$ at this frequency gives the required controller ain.[5,2] The values thus obtained for μ and the period of scillation were 27·4 and 44 sec, respectively.



ig. 12.—Closed-loop transient responses to step-function disturbances.
(a) Proportional control; disturbance in θ_i.
(b) Proportional control; disturbance in T_i.
(c) Three-term control; disturbance in θ_i.
(d) Three-term control; disturbance in T_i.
Calibration pulses at 50 sec intervals.

Fig. 12 shows the transient responses to step-function disturbances in θ_i and T_i which were obtained by means of the analogues and using the above value for μ. The numerical esults are given in Table 3, and in view of the difficulty in neasuring the subsidence ratio accurately, there is good agreenent with the predicted values for the period and damping.

(5.2) Three-Term Control

When the maximum deviation and the offset occurring with proportional control are excessive, the derivative and integral erms in the controller may be employed. The introduction of lerivative control normally increases the frequency of the oscillatory component in the closed-loop transient response, and n the case of the heat exchanger this increase would take the requency outside the accurate range of the analogue. In order o avoid this, two additional lags were introduced into the control loop. This was a realistic approach since measuring and transmission lags frequently occur in practical systems.

Transmission lags arise when the controller (usually pneumatically operated) is situated at some distance from the plant, and they occur between the measuring unit and the controller and/or between the controller and the regulating unit. In this instance only one transmission lag was considered; it was assumed to be in the former position in the loop and to be exponential in nature. Measuring lags are often very large, particularly in the case of temperature-detecting elements which are contained in protective wells. Typical values for the time-constants of measuring and transmission lags are 30 sec and 10 sec, respectively,[6] and these were be no means negligible compared with the lags in the heat exchanger. The electrical analogues of two such exponential lags were introduced into the control loop as shown in Fig. 9 (R_MC_M and R_TC_T).

The complete transfer function for the loop was then formidable indeed but the optimum values for μ, D and I were rapidly determined by means of the analogue. The resulting transient responses are shown in Fig. 12 while the numerical data are given in Table 3. It should be noted that the oscillograms depict the behaviour of T and not the output of the measuring unit. The latter is the only quantity which could be observed in an actual system and it would give a false impression of the quality of control.

(6) CONCLUSIONS

The practicability of a passive-network electrical analogue for a distributed-parameter heat exchanger has been demonstrated, and it has been shown that such an analogue can be usefully employed in order to study the dynamic behaviour of the heat exchanger when in an automatic-control loop.

In one respect—in the approximation to the exponential function—an adequate degree of accuracy can be obtained by using a sufficiently large number of sections. The simulation of the thermal capacity of the tube should not appreciably alter this state of affairs, although a practical limitation arises here in that non-resistive inductors for the series impedances are the ideal requirement. The axial conductivity of the tube might also be allowed for in the analogue, as indicated in Section 2.1, but only at the expense of still further increasing the complexity of the series impedance.

The frequency-dependent errors in the complete analogue which occurred within the accurate bandwidth of the tube analogue appeared to be due to the method of terminating the latter, although the use of the additional five-section unit did not effect as much improvement as had been expected. The feedback analogue does not suffer from this disadvantage, but the greater number of sections required weighs rather heavily against it. However, the relative independence of the adjacent sections in the feedback analogue, which is a fundamental requirement but which, nevertheless, gives rise to the greater

Table 3

Experimental Results for Automatic Control of T

Type of control	Controller settings			Deviation reduction factor		Period of oscillation		Approx. subsidence ratio
	μ	D	I	Disturbance in θ_i	Disturbance in T_i	Predicted	Measured	
		sec	sec			sec	sec	
Proportional ..	27·4	0	∞	18·2	15·5	44	45	3·2
Three-term ..	14·5	30·5	85	7·9	7·8	—	87	3·0

The deviation reduction factor is the ratio of the maximum deviation produced by a step-function disturbance in the absence of control to the maximum deviation occurring when under control.

inaccuracy in the approximation to the exponential function, might be used to advantage. By making the amount of feed-back an appropriate function of the voltage representing fluid temperature it might be possible to simulate the effects of exothermic or endothermic reactions taking place in the tube fluid. Such processes frequently occur in the chemical industry.*

Both types of analogue might be employed to simulate concentric-tube heat exchangers. In such cases the dependence of T upon x would introduce a further complication into the feed-back analogue; a phase-reversing amplifier of unity gain would be required for each section in order to extract $\Delta\theta_r$ and ΔT_r. In addition the number of components per section would be approximately doubled.

Finally, in so far as the practical application of the analogues which have been discussed and proposed is concerned, we must bear in mind their limitations and the assumptions made in their derivation. There must be a sufficient number of sections to give an adequate bandwidth. Without carrying out extensive calculations and tests, the steady-state error, which should be considerably less than the maximum allowable frequency-dependent error, provides some indication of this.

The invariable nature of the method of simulating the velocity of the inner fluid flow in the passive-network analogue constitutes a serious disadvantage since in many control applications it would be easiest to regulate this quantity. However, this does not matter in cases where v_1 must be held constant and control effected by regulating v_2.

Most important of all, we must not lose sight of the fact that the analogues are based on a simplified analysis of the heat-exchange process. In practice the parameters α and w are subject to variation with tube contamination and temperature, respectively. Turbulent flow of the inner fluid and inefficient mixing in the outer vessel are other factors which have not been considered. The validity of the assumptions can only be satisfactorily determined by experiment, and in this connection attention might be drawn to Takahashi's automatic-control experiments with a concentric-tube heat exchanger.[1] He obtained satisfactory results with controller settings predicted by means of equations based on assumptions identical to those which have been used here.

(7) ACKNOWLEDGMENTS

The author is indebted to Professor J. C. Prescott of King's College, Newcastle upon Tyne, for the facilities placed at his disposal during the course of this work, and to Mr. F. J. U. Ritson of the Electrical Engineering Department, and Dr. J. D. Weston of the Mathematics Department, King's College, for their advice and encouragement.

The author's interest in analogues for heat exchangers was aroused by Messrs. A. J. Young, A. R. Aikman and I. Gray of Imperial Chemical Industries Ltd., who also supplied the numerical data on which the design of the tube analogue was based. Mr. Gray's influence on the author's choice of the heat exchangers which were selected for study is also acknowledged.

(8) REFERENCES

(1) TAKAHASHI, Y.: "Transfer Function Analysis of Heat Exchange Processes," from "Automatic and Manual Control" (Butterworth, London, 1952).

(2) FORD, R. L.: "The Determination of Optimum Process-Controller Settings and their Confirmation by Means of an Electronic Simulator," *Proceedings I.E.E.*, Paper No. 1533 M, July, 1953 (**101,** Part II, p. 141).

* See contribution of S. T. Lunt to discussion on Reference 2.

(3) RUTHERFORD, C. I.: "The Practical Application of Frequency Response Analysis to Automatic Control Problems," *Proceedings of the Institution of Mechanical Engineers*, 1950, **162**, p. 334.

(4) AIKMAN, A. R.: "The Frequency Response Approach to Automatic Control Problems," *Transactions of the Society of Instrument Technology*, 1950, **3**, p. 2.

(5) KUSTERS, N. L., and MOORE, W. J. M.: "A Generalisation of the Frequency Response Method for the Study of Feed-back Control Systems," from "Automatic and Manual Control" (Butterworth, London, 1952).

(6) AIKMAN, A. R.: "The Influence of Measuring and Transmission Lags," *ibid.*

(7) LAWSON, D. I., and MCGUIRE, J. H.: "The Solution of Transient Heat Flow Problems by Analogous Electrical Networks," *Proceedings of the Institution of Mechanical Engineers*, A, 1953, **167**, p. 275.

(8) PASCHKIS, V.: "The Heat and Mass Flow Analyser—a Tool for Heat Research," *Metal Progress*, 1947, **52**, p. 813.

(9) MALAVARD, L., and MIROUX, J.: "Electrical Analogies for Heat Transfer Problems," *Engineers Digest*, 1952, **13**, p. 417.

(10) DE LACLÉMANDIÈRE, J.: "Étude experiméntale de la transmission de la chaleur en régime variable à l'aide de la méthode des analogies électrique et thermique," *Chaleur et Industrie*, 1947, **23**, p. 293, and 1948, **24**, p. 14.

(9) APPENDICES

(9.1) Analysis of the Constant-Jacket-Temperature Heat Exchanger

(9.1.1) Heat-Transfer Equations.

Consider an element of the heat-exchanger tube of length δx. If $\theta > T$ the rate of heat transfer from the inner fluid to the tube is

$$\alpha_1(\theta - \theta_T)U\delta x \quad . \quad . \quad . \quad . \quad . \quad (61)$$

Similarly, the rate of heat transfer from the tube to the outer fluid is

$$\alpha_2(\theta_T - T)U\delta x \quad . \quad . \quad . \quad . \quad . \quad (62)$$

The difference between eqns. (61) and (62) gives the rate of storage of heat in the tube element, i.e. $(w_2\delta x)\partial\theta_T/\partial t$; so it follows that

$$w_2\frac{\partial\theta_T}{\partial t} = \alpha_1 U(\theta - \theta_T) - \alpha_2 U(\theta_T - T) \quad . \quad . \quad (9)$$

The total rate of change of the temperature of the inner fluid with respect to time is given by

$$\frac{d\theta}{dt} = \frac{\partial\theta}{\partial t} + \frac{dx}{dt}\frac{\partial\theta}{\partial x}$$

which may be written

$$\frac{d\theta}{dt} = \frac{\partial\theta}{\partial t} + v_1\frac{\partial\theta}{\partial x} \quad . \quad . \quad . \quad . \quad . \quad (63)$$

The total rate of loss of heat from an element of length δx is $-(w_1\delta x)d\theta/dt$, and this must be the rate of heat transfer to the tube element. Hence from eqns. (63) and (61) we have

$$\frac{\partial\theta}{\partial t} + v_1\frac{\partial\theta}{\partial x} = -\frac{\alpha_1 U}{w_1}(\theta - \theta_T) \quad . \quad . \quad . \quad . \quad (8)$$

Eqn. (1) is obtained by eliminating θ_T between eqns. (8) and (9).

.2) Transfer Functions.

f $\bar{\theta}$ and $\bar{T}$ are the Laplace transforms with respect to t of , t) and $T(t)$, respectively, the transform of eqn. (1) is as ows:

$$\bar{\theta} - [s\theta(x,0) - \dot{\theta}(x,0)] + v_1\frac{\partial}{\partial x}[s\bar{\theta} - \theta(x,0)]$$

$$+ b_1[s\bar{\theta} - \theta(x,0)] + v_1 b_2\frac{\partial\bar{\theta}}{\partial x} + b_3\bar{\theta} = b_3\bar{T}$$

hen $\theta(x,0) = \dot{\theta}(x,0) = 0$ this reduces to

$$v_1(s + b_2)\frac{\partial\bar{\theta}}{\partial x} + (s^2 + b_1 s + b_3)\bar{\theta} = b_3\bar{T}$$

egrating with respect to x we obtain

$$\bar{\theta} = \kappa\varepsilon^{-\left(\frac{s^2+b_1s+b_3}{s+b_2}\right)\frac{x}{v_1}} + \frac{b_3\bar{T}}{s^2 + b_1 s + b_3} \quad . \quad . \quad (64)$$

here κ is the constant of integration. If $T = 0$ for all values of and t and if $\bar{\theta}_i$ is the Laplace transform of θ when $x = 0$, n. (64) reduces to

$$\frac{\bar{\theta}}{\bar{\theta}_i} = \varepsilon^{-\left(\frac{s^2+b_1s+b_3}{s+b_2}\right)\frac{x}{v_1}} \quad . \quad . \quad . \quad . \quad . \quad . \quad (65)$$

ibstituting $s = j\omega$ and $x = l$ we obtain the transfer function r a sinusoidal disturbance in θ_i as given by eqn. (3).

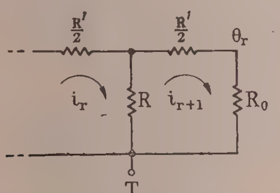
If $\theta_i = 0$ for all values of t then $\bar{\theta} = 0$ when $x = 0$, and in ese circumstances eqn. (64) gives

$$\kappa = \frac{-b_3\bar{T}}{s^2 + b_1 s + b_3}$$

ibstituting this value for κ in eqn. (64) the result is

$$\frac{\bar{\theta}}{\bar{T}} = \frac{b_3}{s^2 + b_1 s + b_3}\left[1 - \varepsilon^{-\left(\frac{s^2+b_1s+b_3}{s+b_2}\right)\frac{x}{v_1}}\right] \quad . \quad . \quad (66)$$

ibstituting $s = j\omega$ and $x = l$ we obtain the transfer function r a sinusoidal disturbance in T as given by eqn. (4).



g. 13.—Equivalent circuit for rth section of perfectly matched analogue.

(9.2) Analysis of First Passive-Network Analogue

2.1) Steady-State Response.

When the analogue is terminated in R_0 the network is perfectly atched and an equivalent circuit for the rth section is shown Fig. 13. By the branched-circuit rule we have

$$i_{r+1} = i_r\frac{R}{R + \frac{R'}{2} + R_0} = \frac{i_r}{1 + \frac{R'}{2R} + \frac{R_0}{R}}$$

follows from this relationship that

$$i_{r+1} = i_1\left(1 + \frac{R'}{2R} + \frac{R_0}{R}\right)^{-r}$$

And since $i_1 = (\theta_i - T)/R_0$ this becomes

$$i_{r+1} = \frac{(\theta_i - T)}{R_0}\left(1 + \frac{R'}{2R} + \frac{R_0}{R}\right)^{-r} \quad . \quad . \quad . \quad (67)$$

We see from Fig. 13 that θ_r is given by

$$\theta_r = i_{r+1}R_0 + T$$

Substituting for i_{r+1} from eqn. (67) we obtain

$$\theta_r = (\theta_i - T)\left(1 + \frac{R'}{2R} + \frac{R_0}{R}\right)^{-r} + T \quad . \quad . \quad (68)$$

Putting $r = n$ and substituting the appropriate values for R', R and R_0 from eqns. (19), (20) and (7) we obtain the steady-state response of the network as given by eqns. (21) and (22).

(9.2.2) Frequency Response.

If θ_i is sinusoidal and $T = 0$, and if we assume that the network is terminated in Z_0, a procedure identical with that given in the last Section gives the following expression for θ_n:

$$\theta_n = \theta_i\left(1 + \frac{Z_1}{2Z_2} + \frac{Z_0}{Z_2}\right)^{-n} \quad . \quad . \quad . \quad . \quad (69)$$

Substituting for Z_2, Z_1 and Z_0 from eqns. (15), (16) and (17) we obtain eqns. (24) and (25).

The response of the analogue to a sinusoidal disturbance in T with $\theta_i = 0$ may be determined with reference to Fig. 14 as

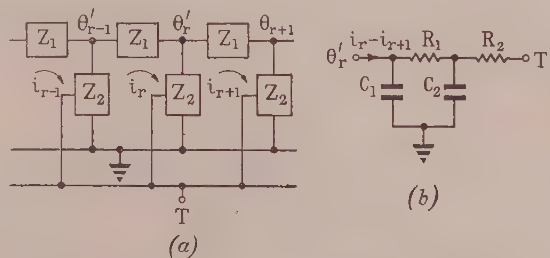


Fig. 14.—(*a*) Schematic of a portion of the analogue.
(*b*) The rth shunt impedance.

follows. An analysis of the circuit of Fig. 14(*b*) shows that the current flowing into the rth Z_2 is given by

$$i_r - i_{r+1} = \frac{\theta'_r}{Z_2} - \frac{T}{Z} \quad . \quad . \quad . \quad . \quad . \quad (70)$$

where Z is given by eqn. (27). Expressing i_r and i_{r+1} in terms of θ'_{r-1}, θ'_r, θ'_{r+1} and Z_1, eqn. (70) becomes

$$-\theta'_{r-1} + \theta'_r\left(2 + \frac{Z_1}{Z_2}\right) - \theta'_{r+1} = \frac{Z_1 T}{Z} \quad . \quad . \quad (71)$$

The general solution of eqn. (71) is $\theta'_r = A\gamma^r + B$ where A and B are arbitrary constants. Substituting this expression for θ'_r in eqn. (71) we obtain

$$-A\gamma^{r-1}\left[\gamma^2 - \left(2 + \frac{Z_1}{Z_2}\right)\gamma + 1\right] + \frac{BZ_1}{Z_2} = \frac{Z_1 T}{Z} \quad . \quad (72)$$

It follows that $B = Z_2 T/Z$ and, since $A\gamma^{r-1} \neq 0$, the value for γ must be such that the quadratic function is zero. This gives two possible values for γ:

$$\gamma_1 = 1 + \frac{Z_1}{2Z_2} + \frac{Z_0}{Z_2} \quad . \quad . \quad . \quad . \quad (73a)$$

$$\gamma_2 = 1 + \frac{Z_1}{2Z_2} - \frac{Z_0}{Z_2} \quad . \quad . \quad . \quad . \quad (73b)$$

In the quadratic function the constant term and the coefficient of γ^2 are both unity so the product of the roots must be unity, i.e. $\gamma_1 = 1/\gamma_2$. We can therefore dispense with the suffixes for γ, and the solution of eqn. (71) becomes

$$\theta'_r = A_1\gamma^r + A_2\gamma^{-r} + \frac{Z_2 T}{Z} \quad . \quad . \quad . \quad . \quad (74)$$

where γ is given by eqn. (73*a*). Since θ_r is the output voltage of the rth T-section we have

$$\theta_r = \tfrac{1}{2}(\theta'_r + \theta'_{r+1}) \quad . \quad . \quad . \quad . \quad . \quad . \quad (75)$$

Substituting for θ'_r and θ'_{r+1} from eqn. (74) we obtain

$$\theta_r = \frac{A_1}{2}(1+\gamma)\gamma^r + \frac{A_2}{2}\left(1+\frac{1}{\gamma}\right)\gamma^{-r} + \frac{Z_2 T}{Z} \quad . \quad (76)$$

Assuming the analogue to be terminated in Z_0 it can be shown that the equation for the nth section is

$$\theta'_{n-1} = \gamma\theta'_n - \left(\frac{Z_1}{Z} + \frac{2Z_1}{Z_1 + 2Z_0}\right)T \quad . \quad . \quad . \quad (77)$$

Substituting for θ'_{n-1} and θ'_n from eqn. (74) we find that A_2 vanishes and the resulting value of A_1 is given by

$$A_1 = -\frac{Z_1(Z - Z_2)T}{2Z_0 Z(1-\gamma)\gamma^n} \quad . \quad . \quad . \quad . \quad (78)$$

We know that $\theta = 0$ when $r = 0$, which enables us to determine the value of A_2 from eqn. (76). The final expression for θ_n is then obtained from eqn. (76) by putting $r = n$ and the result is eqn. (26).

(9.3) Analysis of First Feedback Analogue

If we neglect the series resistances ΔR, eqn. (70), which applies to the circuit of Fig. 14(*b*), may be used in order to analyse the feedback analogue. Instead of $i_r - i_{r+1}$ and θ'_r we may write Δi_r and $\theta_{r-1} - \Delta\theta_r$ ($= \theta_r$), respectively, in this case and the equation becomes

$$\Delta i_r = \frac{\theta_{r-1} - \Delta\theta_r}{Z_2} - \frac{T}{Z} \quad . \quad . \quad . \quad . \quad (79)$$

Now $\Delta\theta_r$ is effectively derived by passing Δi_r through a resistance of value $1/v_1 w_1$ so that the following expression results for $\Delta\theta_r$:

$$\Delta\theta_r = \frac{\Delta i_r}{v_1 w_1} = \frac{1}{v_1 w_1}\left(\frac{\theta_{r-1} - \Delta\theta_r}{Z_2} - \frac{T}{Z}\right) \quad . \quad . \quad (80)$$

Solving eqn. (80) for $\Delta\theta_r$ we obtain

$$\Delta\theta_r = \frac{1}{1 + v_1 w_1 Z_2}\left(\theta_{r-1} - \frac{TZ_2}{Z}\right) \quad . \quad . \quad . \quad (81)$$

From eqn. (81) and the relationship $\Delta\theta_r = \theta_{r-1} - \theta_r$ the following expression for θ_r may be derived:

$$\theta_r = \theta_{r-1}\left(\frac{v_1 w_1 Z_2}{1 + v_1 w_1 Z_2}\right) + \frac{TZ_2}{Z}\left[1 - \left(\frac{v_1 w_1 Z_2}{1 + v_1 w_1 Z_2}\right)\right] \quad . \quad (82)$$

When $T = 0$ it follows directly from eqn. (82) that

$$\theta_r = \theta_i\left(1 + \frac{1}{v_1 w_1 Z_2}\right)^{-r} \quad . \quad . \quad . \quad . \quad (83)$$

Putting $r = n$ and substituting for Z_2 from eqn. (15) the result is eqn. (29), the transfer function for a sinusoidal disturbance in θ_i.

When $T \neq 0$ and $\theta_i = 0$ we find from eqn. (82) that θ_r is given by

$$\theta_r = \frac{TZ_2}{Z}\left[1 - \left(1 + \frac{1}{v_1 w_1 Z_2}\right)^{-r}\right] \quad . \quad . \quad . \quad (84)$$

Again putting $r = n$ and substituting for Z_2 and Z from eqns. (15) and (27) we obtain eqn. (30), the transfer function for a sinusoidal disturbance in T.

The steady-state response of the analogue is obtained from eqns. (83) and (84) by putting $\omega = 0$. The results may be combined by the superposition theorem to give eqn. (28).

(9.4) Transfer Functions for One-Fluid-Mixed Heat Exchanger

If $\bar{\theta}$ and $\bar{T}$ are the Laplace transforms with respect to τ of $\theta(X, \tau)$ and $T(\tau)$, respectively, the transforms of eqns. (31*a*) and (32*a*) are as follows:

$$s\bar{\theta} - \theta(X,0) + \frac{\partial\bar{\theta}}{\partial X} = a_1(\bar{T} - \bar{\theta}) \quad . \quad . \quad . \quad (85)$$

$$\rho[s\bar{T} - T(0)] + (\bar{T} - \bar{T}_i) = a_2\left(\int_0^1 \bar{\theta}\, dX - \bar{T}\right) \quad . \quad (86)$$

When $\theta(X, 0) = T(0) = 0$ these equations reduce to

$$\frac{\partial\bar{\theta}}{\partial X} + (a_1 + s)\bar{\theta} = a_1\bar{T} \quad . \quad . \quad . \quad . \quad (85a)$$

$$(1 + a_2 + \rho s)\bar{T} - \bar{T}_i = a_2\int_0^1 \bar{\theta}\, dX \quad . \quad . \quad . \quad (86a)$$

Integrating eqn. (85*a*) with respect to X we obtain

$$\bar{\theta} = \kappa\varepsilon^{-(a_1+s)X} + \frac{a_1\bar{T}}{a_1 + s} \quad . \quad . \quad . \quad . \quad (87)$$

where κ is the constant of integration. When $X = 0$, $\theta = \theta_i$ for all values of τ so that $\bar{\theta} = \bar{\theta}_i$, and in these circumstances eqn. (87) gives

$$\kappa = \bar{\theta}_i - \frac{a_1\bar{T}}{a_1 + s} \quad . \quad . \quad . \quad . \quad . \quad (88)$$

Substituting this value for κ in eqn. (87) we have

$$\bar{\theta} = \bar{\theta}_i\varepsilon^{-(a_1+s)X} + \frac{a_1\bar{T}}{a_1 + s}[1 - \varepsilon^{-(a_1+s)X}] \quad . \quad . \quad (89)$$

Substituting this expression for $\bar{\theta}$ in eqn. (86*a*), evaluating the integral and solving for $\bar{T}$ we obtain

$$\bar{T} = \frac{a_2(a_1 + s)[1 - \varepsilon^{-(a_1+s)}]\bar{\theta}_i + (a_1 + s)^2\bar{T}_i}{(1 + \rho s)(a_1 + s)^2 + a_2 s(a_1 + s) + a_1 a_2[1 - \varepsilon^{-(a_1+s)}]} \quad . \quad (90)$$

By putting $s = j\Omega$, the transfer functions of eqns. (34) and (36) are obtained from eqn. (90).

Substituting the expression for $\bar{T}$ given by eqn. (90) in eqn. (89) we find that, when $X = 1$,

$$\bar{\theta}_0 = \frac{\{[(1 + \rho s)(a_1 + s)^2 + a_2 s(a_1 + s)]\varepsilon^{-(a_1+s)} + a_1 a_2[1 - \varepsilon^{-(a_1+s)}]\}\bar{\theta}_i + a_1(a_1 + s)[1 - \varepsilon^{-(a_1+s)}]\bar{T}_i}{(1 + \rho s)(a_1 + s)^2 + a_2 s(a_1 + s) + a_1 a_2[1 - \varepsilon^{-(a_1+s)}]} \quad . \quad . \quad (91)$$

g $s = j\Omega$ the transfer functions of eqns. (33) and (35) are ed from eqn. (91).

(9.5) **Derivation of Heat-Transfer Equations**

the circuit of Fig. 6(*b*) we have

$$\Delta i_r = C_1\frac{\partial \theta_r}{\partial t} + \frac{\theta_r - T}{R} \quad . \quad . \quad . \quad . \quad (92)$$

ows from eqn. (12) that the required value of Δi_r is given by

$$\Delta i_r = -v_1 w_1(\theta_r - \theta_{r-1}) \quad . \quad . \quad . \quad . \quad (93)$$

ing the values for Δi_r given by eqns. (92) and (93), we n

$$C_1\frac{\partial \theta_r}{\partial t} + \frac{\theta_r - T}{R} = -v_1 w_1(\theta_r - \theta_{r-1}) \quad . \quad . \quad (94)$$

ituting the values for C_1 and R given by eqns. (39) and (40), (94) reduces to the finite-difference form of eqn. (31).

ain, for the circuit of Fig. 6(*b*) we have,

Current flowing into C_3 and R_3 = Sum of currents supplied the resistors R + Constant current input.

efore

$$C_3\frac{dT}{dt} + \frac{T}{R_3} = \sum_1^n \frac{(\theta_r - T)}{R} + v_2 w_3 T_i \quad . \quad . \quad . \quad (95)$$

ituting the values for C_3, R_3 and R given by eqns. (39), (41) 40) we obtain

$$l\frac{dT}{dt} + v_2(T - T_i) = \frac{U\alpha}{w_3}\sum_1^n (\theta_r - T)\Delta x \quad . \quad . \quad (96)$$

(96) is the finite-difference form of eqn. (32).

the limit, as Δx tends to zero, $(\theta_r - \theta_{r-1})/\Delta x$ may be en as $\partial\theta/\partial x$, and the summation may be written as an al so that eqns. (31) and (32) are obtained. It is only true to say that the latter have been derived from the gue owing to the use of eqn. (93).

(9.6) **Analysis of Second Passive-Network Analogue**

e generalized circuit of Fig. 14(*a*) is also valid for the tube gue of the one-fluid-mixed heat exchanger. In this case Z_2 nposed of R and C_1 and the current flowing into the rth Z_2 en by

$$i_r - i_{r+1} = \theta_r' j\omega C_1 + \frac{\theta_r' - T}{R}$$

efore

$$i_r - i_{r+1} = \frac{\theta_r'}{Z_2} - \frac{T}{R} \quad . \quad . \quad . \quad . \quad . \quad (97)$$

essing i_r and i_{r+1} in terms of θ_{r-1}', θ_r', θ_{r+1}' and Z_1, eqn. (97) be re-arranged to give

$$-\theta_{r-1}' + \left(2 + \frac{Z_1}{Z_2}\right)\theta_r' - \theta_{r+1}' = \frac{Z_1 T}{R} \quad . \quad . \quad (98)$$

(98) is identical in form with eqn. (71) so that its solution be of the same form as eqn. (74) and we may write imme- y that

$$\theta_r' = B_1\lambda^r + B_2\lambda^{-r} + \frac{Z_2 T}{R} \quad . \quad . \quad . \quad . \quad (99)$$

$$\lambda = 1 + \frac{Z_1}{2Z_2} + \frac{Z_0}{Z_2} \quad . \quad . \quad . \quad . \quad (100)$$

B_1 and B_2 are arbitrary constants. If we substitute the values for Z_2, Z_1 and Z_0, given by eqns. (42), (43) and (44), eqn. (100) takes the form expressed by eqn. (52).

If we assume that the tube analogue is perfectly matched at its termination, it may be regarded as having an infinite number of sections and it follows that $B_1 = 0$. In these circumstances eqn. (99) reduces to

$$\theta_r' = B_2\lambda^{-r} + \frac{Z_2 T}{R} \quad . \quad . \quad . \quad . \quad . \quad (101)$$

Since θ_r, the output of the rth T-section, is given by $(\theta_r' + \theta_{r+1}')/2$ we may use eqn. (101) to obtain the following expression for θ_r:

$$\theta_r = \frac{B_2}{2}\left(1 + \frac{1}{\lambda}\right)\lambda^{-r} + \frac{Z_2 T}{R} \quad . \quad . \quad . \quad (102)$$

If when $r = 0$, $\theta = \theta_i$ we find that

$$B_2 = \frac{2(\theta_i - TZ_2/R)}{(1 + \lambda^{-1})} \quad . \quad . \quad . \quad . \quad (103)$$

and eqn. (102) may be written

$$\theta_r = \theta_i\lambda^{-r} + \frac{TZ_2}{R}(1 - \lambda^{-r}) \quad . \quad . \quad . \quad (104)$$

Alternatively, if when $r = 0$, $\theta = 0$ then

$$B_2 = -\frac{2TZ_2}{R(1 + \lambda^{-1})} \quad . \quad . \quad . \quad . \quad (105)$$

and eqn. (102) becomes

$$\theta_r = \frac{TZ_2}{R}(1 - \lambda^{-r}) \quad . \quad . \quad . \quad . \quad . \quad (106)$$

In order to determine the responses of the analogue to a sinusoidal disturbance in θ_i we may take $T_i = 0$, and in this circumstance we have

$$T = Z_3\sum_1^n \frac{(\theta_r' - T)}{R} \quad . \quad . \quad . \quad . \quad (107)$$

where Z_3 is the impedance of C_3 and R_3 in parallel. Substituting for θ_r' from eqn. (101) we obtain

$$T = \frac{Z_3}{R}\sum_1^n \left(B_2\lambda^{-r} + \frac{TZ_2}{R}\right) - \frac{nZ_3 T}{R} \quad . \quad . \quad (108)$$

The evaluation of the summation gives

$$T = \frac{Z_3 B_2}{R}\left(\frac{1 - \lambda^{-n}}{\lambda - 1}\right) + \frac{nZ_3 T}{R}\left(\frac{Z_2}{R} - 1\right) \quad . \quad . \quad (109)$$

Substituting for B_2 from eqn. (103) and solving for T we obtain

$$\frac{T}{\theta_i} = \frac{R(1 - \lambda^{-n})/n}{Z_0\left[\frac{R^2}{nZ_2 Z_3} + \left(\frac{R}{Z_2} - 1\right)\right] + \frac{Z_2}{n}(1 - \lambda^{-n})} \quad . \quad (110)$$

Substituting the appropriate values for R, Z_2 and Z_3, eqn. (110) reduces to the transfer function of eqn. (49). And if the expression for T which is given by eqn. (49) be substituted in eqn. (104) we obtain, when $r = n$, the transfer function of eqn. (48).

For a sinusoidal disturbance in T_i we may take $\theta_i = 0$ so that eqns. (105) and (106) apply in this case. If i_T is the total current flowing through the resistances R we have

$$i_T = \sum_1^n \frac{(T - \theta_r')}{R} \quad . \quad . \quad . \quad . \quad . \quad (111)$$

Substituting for θ_r' from eqn. (101) and evaluating the summation, using the value for B_2 given by eqn. (105), we obtain

$$i_T = \frac{TZ_2^2}{R^2Z_0}(1 - \lambda^{-n}) + \frac{nT}{R}\left(1 - \frac{Z_2}{R}\right) \quad . \quad . \quad (112)$$

The constant current fed to the outer-vessel analogue is $v_2w_3T_i$, so that the current flowing into Z_3 is $(v_2w_3T_i - i_T)$. It follows that

$$T = Z_3(v_2w_3T_i - i_T) \quad . \quad . \quad . \quad . \quad . \quad . \quad (113)$$

Substituting for i_T from eqn. (112), the solution of eqn. (113) for T gives

$$\frac{T}{T_i} = \frac{Z_0v_2w_3R^2/nZ_2}{Z_0\left[\dfrac{R^2}{nZ_2Z_3} + \left(\dfrac{R}{Z_2} - 1\right)\right] + \dfrac{Z_2}{n}(1 - \lambda^{-n})} \quad . \quad (114)$$

The transfer function of eqn. (51) is obtained from eqn. (114) when the appropriate values for R, Z_2 and Z_3 are employed. Finally, the θ_n/T_i response may be derived by substituting in eqn. (106) the value for T given by eqn. (51).

The steady-state responses of the analogue are obtained from eqns. (48)–(52) by putting $\Omega = 0$. In this case no assumptions are involved since the tube analogue is correctly terminated in R_0.

(9.7) Analysis of Second Feedback Analogue

Neglecting the effects of the series resistances ΔR, the equation for Δi_r for the feedback analogue can be obtained directly from eqn. (97) and we may write

$$\Delta i_r = \frac{\theta_{r-1} - \Delta\theta_r}{Z_2} - \frac{T}{R} \quad . \quad . \quad . \quad . \quad (115)$$

Eqn. (115) is identical in form to eqn. (79) so that the solution for θ_r is of the same form as eqn. (82). We have, therefore,

$$\theta_r = \theta_{r-1}\left(\frac{v_1w_1Z_2}{1 + v_1w_1Z_2}\right) + \frac{TZ_2}{R}\left[1 - \left(\frac{v_1w_1Z_2}{1 + v_1w_1Z_2}\right)\right] \quad . \quad (116)$$

If when $r = 1$, $\theta_{r-1} = \theta_i$ then eqn. (116) may be written

$$\theta_r = \theta_i\lambda_1^{-r} + \frac{TZ_2}{R}(1 - \lambda_1^{-r}) \quad . \quad . \quad . \quad (117)$$

where λ_1 is given by eqn. (55). And if $T_i = 0$ then

$$T = Z_3\sum_1^n \frac{(\theta_r - T)}{R} \quad . \quad . \quad . \quad . \quad (118)$$

Substituting for θ_r from eqn. (117) and evaluating the summation we obtain

$$T = \frac{v_1w_1Z_2Z_3}{R}\left(\theta_i - \frac{Z_2T}{R}\right)(1 - \lambda_1^{-n}) + \frac{nZ_3T}{R}\left(\frac{Z_2}{R} - 1\right) \quad . \quad (119)$$

Solving eqn. (119) for T the result is

$$\frac{T}{\theta_i} = \frac{v_1w_1R(1 - \lambda_1^{-n})/n}{\dfrac{R^2}{nZ_2Z_3} + \left(\dfrac{R}{Z_2} - 1\right) + \dfrac{v_1w_1Z_2}{n}(1 - \lambda_1^{-n})} \quad . \quad (12$$

Substituting the appropriate values for R, Z_2 and Z_3 in eqn. (12 we obtain the transfer function of eqn. (54). And substitutir the value for T given by eqn. (54) in eqn. (117) the result, whe $r = n$, is the transfer function of eqn. (53).

(9.8) Scale Factors and Numerical Data

Let 1 volt $\equiv p$°C
1 millicoulomb $\equiv q$ calories
1 second $\equiv m$ seconds

Then 1 milliampere $\equiv \dfrac{q}{m}$ cal/sec

Now

$$\text{Thermal resistance} = \frac{\text{Temperature difference}}{\text{Heat flow}}$$

Therefore, 1 ohm $\equiv \dfrac{pm}{q} \times 10^{-3}$ °C/cal/sec

Also,

$$\text{Thermal capacity} = \frac{\text{Heat stored}}{\text{Temperature}}$$

Therefore, 1 farad $\equiv \dfrac{q}{p} \times 10^{3}$ cal/°C

It follows that

(*a*) Resistance: 1° C/cal/sec $\equiv \dfrac{q}{pm}$ kilohms

(*b*) Capacitance: 1 cal/°C $\equiv \dfrac{p}{q} \times 10^{3}\,\mu$F

The numerical values used were $m = 10^5$ and $q/p = 4 \times 1$(so that

1° C/cal/sec $\equiv$ 40 kilohms
1 cal/° C $\equiv 2{\cdot}5 \times 10^{-4}\,\mu$F

It is unnecessary to assign specific values to p and q wh dealing only with temperature and voltage ratios.

The values used for the heat-exchanger parameters were follows, the figures in brackets being the original parameters:

l = 500 (635) cm.
$U\alpha$ = 0·219 (0·183) cal/cm sec °C.
w_1 = 4 (3·85) cal/cm °C.
w_3 = 40 cal/cm °C.
v_1 = 20·4 (18·1) cm/sec.
v_2 = 2·04 cm/sec.

.3.42 : 621.311.6 : 621.317.7.089.6

Paper No. 1937 M
Jan. 1956

ELECTRONIC SUPPLY FOR USE IN THE CALIBRATION OF INSTRUMENTS

By F. J. WILKINS, B.A., B.Sc., and S. HARKNESS, Graduate.

(*The paper was first received* 12*th October*, 1954, *and in revised form* 2*nd August*, 1955.)

SUMMARY

rief review is given of some of the features of supplies that are le for use in the calibration of a.c. instruments, and an account n given of a high-power oscillator-amplifier set that has been or this purpose.

set consists of an oscillator, two amplifiers and a phase-shift An output of at least 700 VA is available at unity power factor the range 30 c/s–5 kc/s from each amplifier, and if necessary the utputs can be connected in series. The output voltage is free random variations greater than 0·01%, drifts at maximum power t by not more than 0·003% per minute, and has a supply-fre- y component not greater than 0·01%. The output voltage is ore always constant to 0·01% for the time taken to calibrate a on an instrument scale.

distortion in the outputs of the oscillator-amplifier set is never r than 0·45% of the fundamental at any frequency or power t within its rating.

(1) INTRODUCTION

finitions.—The term "stability" is used only in the sense an amplifier is stable if it satisfies the Nyquist criterion.

e term "voltage constancy" is used to indicate the steadiness e output voltage.

laboratory intended for the testing of a.c. instruments must rovided with reference standards, transfer or comparison ments and suitable supplies. Whilst each will set a limit e accuracy that can be achieved, the behaviour of the transfer ment and the supply also affect the speed and ease with the results can be taken. The performance of the supply refore important, and for the rapid and accurate testing of grade instruments it is found desirable that the voltage d be constant to 0·01% for a period of approximately a te, and that the harmonic content should be as low as ble and preferably not greater than 0·5% of the fundamental. plies having these qualities may be realized in three ways, ly

) Alternators driven by d.c. motors supplied from a battery.
) Alternators provided with some form of stabilizing equipment, en by motors supplied from the mains.
) Electronic supplies.

h has its inherent advantages and disadvantages, and the e of the system to be used will depend on the frequency , power requirements, and such factors as the space available use the supply and the permissible capital expenditure.

e alternator is the more direct of the two systems and the economical for frequencies up to a few hundred cycles cond when power outputs greater than about 25 kVA are ed. For a supply of this magnitude the initial capital , even if the motor driving the alternator is supplied by cid storage batteries, would almost certainly be smaller hat of the corresponding electronic supply, and the design aintenance would be less complicated. The size of the ator itself would also compare favourably with that of the onic equipment, although the use of batteries to give the necessary output-voltage constancy would add considerably to the total space required. Batteries have other disadvantages: their initial cost and the cost of their maintenance are both high, and they must be charged. Thus, the most compact and economic form of the alternator supply is that which incorporates some system which stabilizes both frequency and output voltage, thereby allowing the motors to be driven from the mains.

Further considerations of the merits of the two systems show that they are in many ways complementary. Whilst the performance of the alternator is satisfactory for frequencies below a few hundred cycles per second, it is unsuitable at higher frequencies because of the large amount of distortion that occurs. An electronic supply, on the other hand, can give a good waveform over a wide frequency band and is normally used when outputs smaller than 100 VA are required. For power levels of this magnitude the electronic supply is cheaper and more compact, and it also has the advantage that the frequency is constant with time, independent of load conditions, and may be varied easily.

The high-power oscillator-amplifier set described in the paper gives two outputs, each of at least 700 watts, and it is estimated that the initial cost is approximately one-fifth to one-tenth that of an alternator system of comparable output. It is comparatively small, is silent in operation, but has a slight disadvantage in that the energy dissipated as heat is concentrated and must be suitably dispersed. The installation has now been in operation for approximately a year, and in this period maintenance problems have been few. It is early, however, to make any comparison with alternators on this point.

Both alternators driven by d.c. motors supplied from batteries and oscillator-amplifier sets are used at the N.P.L., and details of these supplies have been given in a recent paper.[1] The alternators are used mainly for frequencies below 100 c/s, although one will give a supply up to 500 c/s, whilst the electronic supply, which was installed before the development of the supply to be described was undertaken, covered the frequency range 300 c/s–100 kc/s. Experience has shown that both systems can give the required voltage constancy, although in practice the output from the alternators is the less stable of the two, and variations greater than 0·02% may occur owing to the variable resistance between the motor brushes and commutator. On the other hand, whilst a double output of 30 kVA is possible from one alternator, the maximum power available from this first electronic supply under optimum conditions is 320 VA. The high-power oscillator-amplifier set bridges the gap between the existing supplies, giving an electronic supply at 50 c/s and a supply of greater power output at 5 kc/s.

For most of the a.c. measurements made at the N.P.L. a power of about 1 kW is adequate, and this represents the maximum power that can be conveniently developed from a pair of radiation-cooled triodes. Commercial public-address amplifiers of this output have been available for many years, but so far as is known there has not been available as a unit of an electronic generator an amplifier of this output which would meet the exacting requirements of a supply for accurate instrument calibration. The paper does not present any new principles

en contributions on papers published without being read at meetings are or consideration with a view to publication.
aper is an official communication from the National Physical Laboratory.

but shows that by careful design it is possible to construct such an electronic generator giving outputs of at least 700 watts in the frequency range 30 c/s–5 kc/s. It consists of an oscillator, two amplifiers and a phase-shift unit. Each will be discussed separately.

(2) OSCILLATOR

The first electronic supply used at the N.P.L. for the testing of a.c. instruments in the Electrotechnics Section was based on an *RC* oscillator and a feedback amplifier designed at the Post Office Research Station. The oscillator developed to drive the high-power amplifiers is a modified form of the Post Office design and covers the range 30 c/s–30 kc/s, supplying 8 mW into a 600-ohm load.

The basic circuit is shown in Fig. 1, and when the frequency-control networks are disconnected is a 3-stage amplifier having

range, providing the current is maintained at a constant val the temperature coefficient of resistance of the lamps might expected to be constant and of the same order as for tungst The measured coefficient was 0·35% per deg C. Some syst of compensation for temperature changes is therefore desirab and it can be shown that if both halves of the divider have same temperature coefficient the current in the output stage independent of the ambient temperature. The output w therefore be constant, provided that the anode load is consta

This condition for temperature compensation is satisfied making R_{17} of platinum wire having a temperature coefficient 0·35% per deg C, and mounting it as near the lamps as possib Good compensation is given, for although drifts as large −0·07% and +0·05% per hour have been observed, changes in general do not exceed ±0·02% per hour over peric of 3–4 hours' operation.

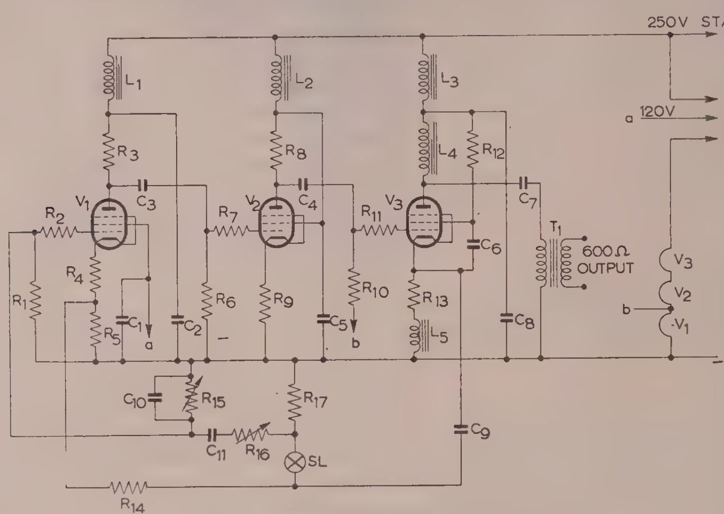


Fig. 1.—Basic oscillator circuit.

R_1 10 MΩ	R_9 470 Ω	C_1, C_2 200 μF	V_1, V_2 UF42
R_2 330 Ω	R_{10} 470 kΩ	C_3, C_4 2 μF	V_3 UL41
R_3 22 kΩ	R_{11} 330 Ω	C_5–C_8 8 μF	
R_4 120 Ω	R_{12} 68 kΩ	C_9 16 μF	
R_5 180 Ω	R_{13} 820 Ω	C_{10}, C_{11} X1, 0·1584 μF; X10, 0·01584 μF	T1, ×1. P.O. T1557; ×10 P.O. A.1194
R_6 1 MΩ	R_{14} 1·9 kΩ		
R_7 330 Ω	R_{15}, R_{16} Wien bridge decades	L_1–L_3 P.O. 3406	S.L. 8—No. 2, 6 V Switchboard lamps
R_8 22 kΩ	R_{17} 84 Ω, platinum (see text)	L_4 P.O. L.1390; L_5 P.O. L.1389	

an overall gain of 21 dB, in which 36 dB of negative feedback is developed in the working range. The feedback loops are returned from the cathode of the final valve and the oscillator output is taken from the anode.

The *RC* frequency control is a Wien bridge network in which the resistive elements are switched to give a decade system and the capacitances to give two ranges. A fine-control variable resistor is also included to make the frequency selection continuous over both ranges.

The amplitude control,[2] which is effectively the cathode load, consists of a number of selected tungsten-filament lamps connected in series with a resistor R_{17} to form a divider from which the Wien bridge network is fed. The ratio of this divider is determined by the current flowing through the output stage. As the time-constant of the lamp assembly is very long compared with the lowest frequency generated, no appreciable harmonic distortion is introduced by the non-linearity of this control.

The current passing through the lamps is much less than the maximum rated value, and the temperature of the filament is therefore comparatively low; hence, over a limited temperature

The oscillator and its stabilized power supply are connect as an integral unit in which series-heater valves are us throughout, the heaters being connected across the stabiliz output. The stabilized power supply is designed on a variati due to Mezger[3] of the conventional series-parallel[4,5] arran ment, and is of the form in which signals from the unstabiliz and stabilized supplies are fed to the screen and grid respectiv of the control valve. The series valve is shunted by a resis through which two-thirds of the current passes. This resis decreases the stabilization range of the system, but it grea increases the magnitude of the current that can be stabilized a also, by allowing current to be fed to the heater chain, enab the valves to heat up when the equipment is first switched on

The output voltage of the stabilized supply is constant to 0·01 and free from surges when the supply voltage is varied over range 210–240 volts, and during periods of eight hours seldo changes by as much as 0·01%. As these performance figu apply to the oscillator heater as well as the h.t. supply, there no short-term unsteadiness resulting from supply variations.

When an output-voltage constancy of better than 0·01%

…ed near the supply frequency and its harmonics, the residual …on the supply, which is less than 300 μV, is still large com… with the alternating voltage levels in the oscillator. Further …ion of the ripple is possible by improving the stabilizing …, but it is considered easier and more efficient to provide …onal filtering in the oscillator. This must be most effective … anode and screen supplies to the first stage. As a further …ution against the introduction of a supply-frequency com…t, the oscillator, which is separated from the power supply, …ased in a Mumetal box. It is estimated that with these …utions, when the oscillator is set to give either a 50 or a …s signal, the supply-frequency component represents less …0·002% of the output.

(2.1) Performance of the Oscillator

… *Dependence on the Supply Voltage.*—When a 6% change is … in the supply voltage within the stabilizing range of the …r supply, 210–240 volts, the corresponding change in the …ator output is less than 0·01%. The drift seldom exceeds …% per hour.

… *Supply-Frequency Component.*—When the oscillator is … a frequency near that of the supply or its harmonics, the …in the output does not exceed 0·002%.

… *Harmonic Content.*—When terminated by a 600-ohm …ive load the harmonic content is not greater than 0·15% of …undamental at frequencies between 50 c/s and 10 kc/s.

(3) AMPLIFIER

It was evident that considerable development would be necessary to produce an amplifier suitable for use in the calibration of instruments, and in order to avoid much tedious work, it was decided to start with a commercial product in which the switching arrangements, the h.t. supplies, the output valves and transformers existed and were suitably housed. A commercial public-address amplifier on which a number of experiments had been made was bought for this purpose, and the required performance was achieved by designing a new amplifier incorporating the existing output valves, the coupling between the driver and the output stage, and the power transformers and rectifying circuits.

(3.1) General Description

The amplifier circuit is as shown in Fig. 2 and consists of five push-pull stages: the output stage is operated in class AB_2. As each stage is symmetrical, reference will be made to one side only in the ensuing considerations.

Voltage feedback is taken from the primary of the output transformer to the grid of the first stage, and some 34–36 dB of harmonic feedback is developed in the frequency range 30 c/s–5 kc/s. Two subsidiary feedback voltages are also introduced in the first stage: one is due to the cathode resistor of the first valve, which is not by-passed, and the other is taken from the anode of the second stage. A further feedback loop of

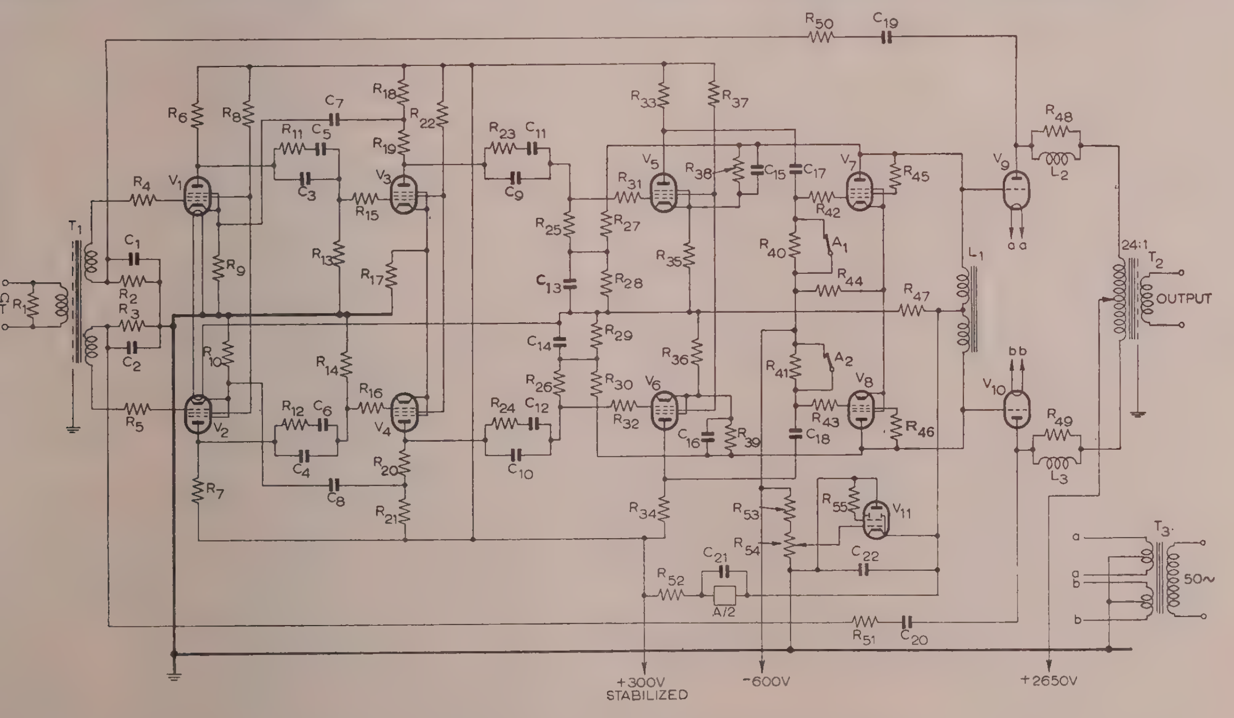


Fig. 2.—Basic amplifier circuit.

R_1 600 Ω	R_{25}, R_{26} 1 MΩ	R_{50}, R_{51} 500 kΩ*	C_{19}, C_{20} 0·5 μF
R_2, R_3 390 Ω*	R_{27}, R_{30} 1 MΩ	R_{52} 220 kΩ	C_{21}, C_{22} 200 μF
R_4, R_5 220 Ω	R_{28}, R_{29} 820 kΩ	R_{53} 82 kΩ	
R_6, R_7 100 kΩ	R_{31}, R_{32} 2·2 kΩ	R_{54} 20 kΩ	V_1, V_2, V_5, V_6 CV 2135
R_8 330 kΩ	R_{33}, R_{34} 100 kΩ		V_3, V_4 CV 138
R_9, R_{10} 2·2 kΩ	R_{35}, R_{36} 3·9 kΩ	C_1, C_2 1 200 μμF	V_7, V_8 EL 37
R_{11}, R_{12} 10 MΩ	R_{37} 330 kΩ	C_3, C_4 0·005 μF	V_9, V_{10} V 1505
R_{13}, R_{14} 1 MΩ	R_{38}, R_{39} 47 kΩ	C_5, C_6 0·1 μF	V_{11} KT 61
R_{15}, R_{16} 220 Ω	R_{40}, R_{41} 220 kΩ	C_7, C_8 68 μμF	T_1 P.O. T.1350 A
R_{17} 330 Ω	R_{42}, R_{43} 2·2 kΩ	C_9, C_{10} 0·25 μF	T_2 Output transformer
R_{18}, R_{21} 1 000 Ω	R_{44} 200 Ω	C_{11}, C_{12} 0·01 μF	L_1 Grid coupling choke
R_{19}, R_{20} 68 kΩ	R_{45}, R_{46} 47 Ω	C_{13}, C_{14} 1·0 μF	L_2, L_3 15 mH
R_{22} 150 kΩ	R_{47} 1 350 Ω	C_{15}, C_{16} 1 000 μμF	
R_{23}, R_{24} 3·3 MΩ	R_{48}, R_{49} 5 kΩ	C_{17}, C_{18} 0·5 μF	A/2 P.O. Relay 50·000 Ω

* Evanohm on 0·01 in mica card.

approximately 36 dB is developed across the third and fourth stages.

The input and output transformers lie outside the feedback loops and ultimately set a limit to the performance of the amplifier; it is therefore necessary that they should introduce the minimum of distortion and voltage variation. Whilst it is equally important to keep the distortion and voltage variations as small as possible in the remainder of the amplifier, they are greatly reduced by the feedback loops, and the voltage constancy of this part of the system is determined by the quality of the feedback resistors.

The amplifier is driven by three conventional power supplies, one stabilized and two unstabilized, having low hum levels to ensure that the supply-frequency component in the output cannot give rise to a beat greater than 0·01%.

(3.2) Output Stage

(3.2.1) Class AB_2 Operation.

An amplifier designed to give a large power output should, for reasons of economy, be made as efficient as possible, and this entails driving the output valves in class B_2. On the other hand, there is the conflicting requirement that the supply should have a low harmonic content and the output valves must not be operated in such a way as to give rise to excessive distortion. It is found that when a compromise is made and the valves are operated in class AB_2 an overall efficiency of approximately 30–40% is possible at maximum output, and yet the distortion when reduced by feedback is within the acceptable limit.

Whilst the amplifier has a relatively high efficiency when the load is resistive, it has the disadvantage that the apparent power output for the maximum rated anode dissipation falls as the power factor of the load is decreased and has dropped, at 50 c/s, from 900 VA at unity power factor to approximately 250 VA at zero power factor. This decrease in the available apparent power, although undesirable, seldom restricts the use of the amplifier in the calibration of instruments, for, since wattmeters are usually tested by the artificial-load method, resistive loads are nearly always used. In addition, any circuit having a power factor less than unity tends to be inductive, and it is then always possible to obtain some compensation by the parallel connection of a suitable capacitor.

One further result of class AB_2 operation is that at frequencies near that of the supply the supply-frequency components developed in the driver stages, and arising from the h.t. supply for the output stage, no longer cancel in the output transformer. The h.t. supply, which has an internal resistance of 300 ohms and gives an output voltage of 2 650 volts at maximum power output, is therefore heavily smoothed, and the residual ripple is reduced to 1·5 volts r.m.s.

The condenser C_{22} reduces the anode dissipation in the output valves by by-passing any out-of-balance or double-frequency currents that would otherwise develop voltages across R_{47}.

(3.2.2) Output Transformer.

The design of transformers used in amplifiers operated in class AB or B has been studied experimentally by Lord,[6] who shows that the harmonic distortion reaches a maximum at approximately one-half the resonant frequency of the circuit and transformer and that this maximum value is related to the ratio L_1/L_2, where L_1 is the leakage inductance between the two halves of the primary and L_2 is the leakage inductance between one-half of the primary and the secondary. The smaller this ratio the smaller is the maximum harmonic distortion. In the present output transformer the ratio is 0·6, the resonant frequency is 80 kc/s and it introduces no appreciable distortion up to 5 kc/s.

The resistive component of the effective output impedance of the amplifier is almost entirely due to the transformer windings whilst the reactive component is composed of the leakage inductance and the inductance L_2. The output resistance is directly in series with the load and, since it has a high temperature coefficient, gives rise to a slow drift in the output, because of self-heating. The rate of drift at maximum power output does not exceed 0·003% per minute.

The primary inductance of the transformer, which has a core of grain-oriented silicon iron, does not vary greatly on drive, the value of each half primary section at full excitation being 100 H.

The transformer is oil-immersed and the insulation between the high-voltage windings and the screens is of polythene, the low permittivity of which helps to reduce the primary capacitance to earth.

The secondary is wound in four sections, so that maximum power output can be developed at 28·75, 57·5 or 115 volts.

(3.3) Driver Stages

The requirements of a driver stage to be used with an output stage in which the valves are driven into grid current have been the subject of much discussion,[7,8] more especially in the years 1934–39 when this particular type of output stage was first used extensively, and it is established that the intermittent flow of grid current causes a variable load on the driver and that the voltage regulation of this stage must be good in order to prevent both amplitude distortion and transient oscillations.

Although the cathode-follower is perhaps the most widely used driver stage, the circuit as shown in Fig. 2 is preferred. It consists of two stages, the third and fourth in the amplifier, across which some 36 dB of feedback is developed, giving a system having a gain of 2 and an output impedance lower than that of the average cathode-follower. The feedback resistors R_{35} and R_{38} also act as a constant load across the choke and greatly reduce the effects of its resonances.

The capacitor C_{15} increases the feedback at high frequencies and enables the high-frequency response of the main feedback loop to be controlled.

The use of the grid coupling choke and the design of the fourth stage to operate between earth and −600 volts enables the last two stages to be directly coupled and is a design feature of the original equipment.

The −600 volt h.t. supply for this stage is also heavily smoothed and the residual ripple is 4 mV r.m.s. When the supply voltage changes, the output from this h.t. supply and that from the 2 650-volt h.t. supply vary in such a way that the bias of the output valves is maintained constant. If, therefore, one of these supplies were stabilized it would also be necessary to stabilize the other.

The flow of grid current through the resistor R_{47} and the heaters of valves V_1 and V_2, which occurs when the amplifier is operated in class AB_2, would normally cause the bias voltage of the output valves to change. Compensation for this current is effected by a method which has been described[9] and consists in connecting a valve V_{11}, which acts as a variable impedance in parallel with R_{47} and the valve heaters. The operating conditions of this valve are adjusted so that the current through it increases when grid current is flowing.

The divider consisting of R_{27} and R_{28} compensates for the fraction of the anode voltage of V_7 developed across R_{35}. C by-passes R_{28} in the working band and influences the amplifier response at low frequencies.

(3.4) Amplifying Stages

tabilized power supply having a ripple of 500 μV r.m.s. the first three stages of the amplifier. Its use avoids the for decoupling networks between these stages, which otherwise be necessary because of the lack of symmetry ch push-pull pair. The low impedance also enables the resistors in each of the first three stages to be combined, by avoiding the need for by-pass capacitors and providing a n which helps to maintain the balance of the two sides.

a further precaution against the introduction of hum the rs of the valves in the first stage are fed from a d.c. source. s requiring a heater current of 150 mA are used, and the rs are connected in series in the earth return of the third ourth stages. As the current is obtained by combining the nt from pairs of valves in push-pull the resultant a.c. voltage all and of signal frequency, and it is not necessary to by- the heaters.

e resistors forming the anode loads of the second stage are ted to have values within 1% of each other, so that equal ls are fed to the driver stages.

e networks comprising R_{11} and C_5, and R_{23} and C_{11} con- d across C_3 and C_9 respectively, are phase-advance networks are necessary to give stability at low frequencies.

(3.5) Input Transformer

the input transformer is loaded, as is usual, the variation of effective winding resistance due to temperature causes a variation in the voltage appearing across the loading resistor. This is analogous to the behaviour of the output transformer, and is avoided by connecting the oscillator load, which is made of Constantan having a low temperature coefficient, on the primary side.

The input transformer has screens between windings and an external Mumetal screen. The screening between windings is essential to prevent the interaction between the two amplifiers that otherwise occurs when a linked or coupled output is developed.

At the lower end of the frequency range the transformer begins to take sufficient magnetizing current to distort the oscillator output, and at the lowest frequencies this represents the major source of distortion in the complete system.

(3.6) Switching System

When the amplifier is switched on, two delay periods are introduced. In the first, which lasts approximately one minute, all heater supplies and the h.t. supply to the fourth stage are connected. The current through the driver stage is therefore able to build up so that the output valves are correctly biased, and the mercury-vapour rectifiers are given time to heat up. The remaining h.t. supplies are then switched on and after a further 15 sec the relay A/2 opens. This breaks the short-circuit on the input to the fourth stage, which is necessary to overcome the effects of transients produced in the system when steady quiescent conditions are being established.

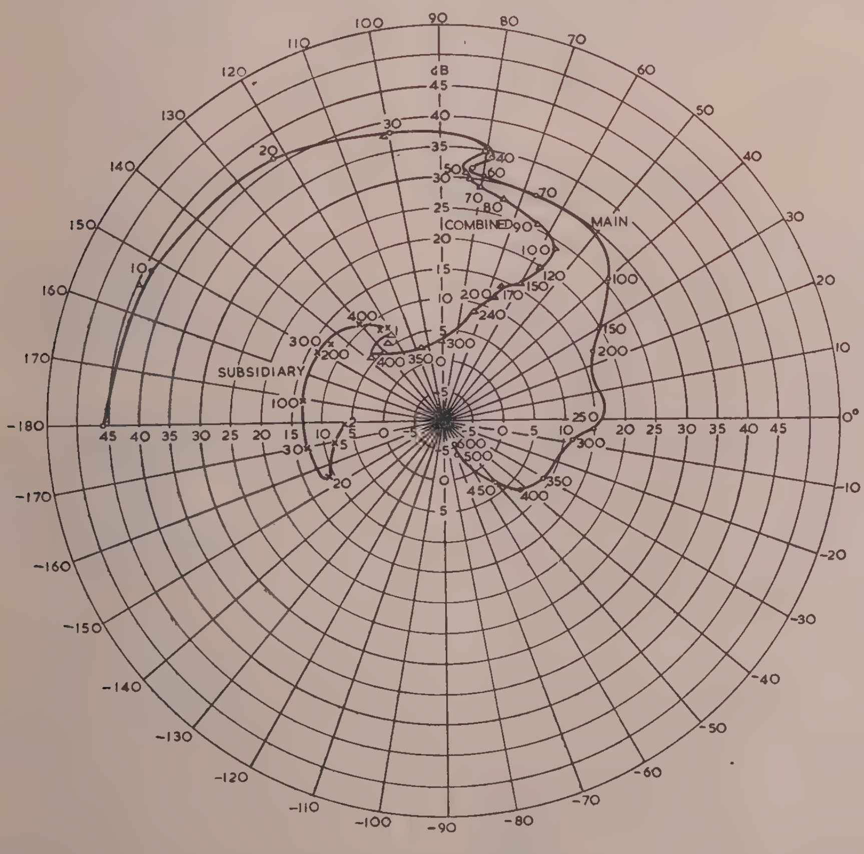


Fig. 3.—Nyquist diagram for high frequencies.

Amplifier terminated with 14 ohms.
Frequency in kilocycles per second. Results by direct measurement.

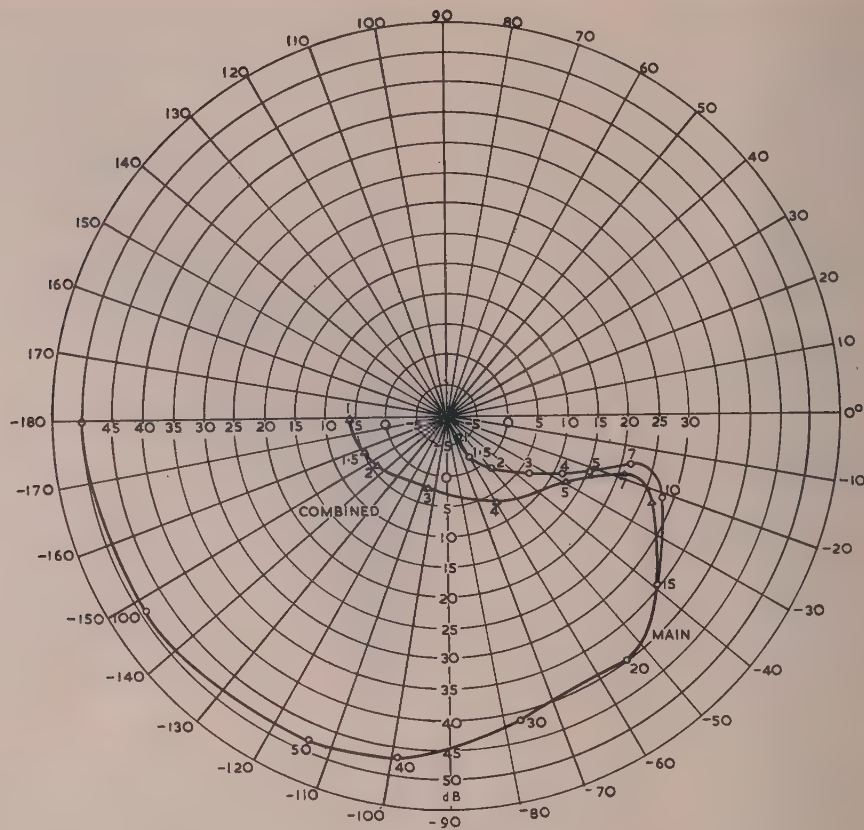


Fig. 4.—Nyquist diagram for low frequencies.

Amplifier termination equivalent to 15 H on each half of transformer primary. Frequency in cycles per second. Results by measurement and calculation.

(3.7) The Feedback System

Three loops are returned to the input, namely

(*a*) The main feedback taken from the anode of the output valve V_9, and inserted in series with one half of the input transformer.

(*b*) The local feedback developed across R_2 in the cathode of V_1, which is operative throughout the whole frequency band and gives a constant loss in harmonic feedback of 10 dB which must be taken into account when the Nyquist diagrams are considered.

(*c*) The feedback introduced into the cathode of V_1 from a tapping on the anode of V_3. This is operative only above the working band.

The return of the feedback from the primary of the output transformer makes the requirement that the amplifier should be stable under all conditions of load easier to satisfy, and an earth-free output, which is a further requirement, is also given.

The feedback loops are designed using methods that are fully described[10, 11, 12] and discussion is limited to a few specific points.

(3.7.1) Stability above the Working Band.

As the gain of a triode is independent of the modulus and argument of the load impedance, providing the value of this impedance is much greater than that of the valve, the inclusion in the anode circuits of R_{48}, which is by-passed in the working range by L_2, makes the gain of the output stage almost independent of termination above 50 kc/s and approximately equal to the amplification factor of the valve. It is also independent of the transformer characteristics, and the effects of resonances, which are difficult to avoid in transformers of this size and rating, do not appear in the main feedback loop.

Fig. 3 shows the resultant feedback locus when the combine subsidiary feedbacks (locus marked "subsidiary") are added t the main feedback loop. The consequent combined feedbac has a large margin of stability and will not become unstab unless the main feedback is increased by another 5 dB without change in the subsidiary loop.

(3.7.2) Stability below the Working Band.

When the amplifier load impedance is infinite the equivale circuit of the output stage is a corrective network of a type th has been discussed.[13] The frequency characteristics of this ne work, providing the generator impedance and the prima winding resistance are regarded as constant, are controlled b the inductance, and the time-constants of the two phase-advan networks and the feedback circuit are so adjusted that the on limitation on stability is a maximum value of this inductanc This maximum value is made greater than the primary inductan of the output transformer.

A condition of minimum stability occurs with an effecti primary inductance of 15 H and this is shown by the cur marked "main" in Fig. 4. A similar curve is shown in Fig. 5 f the condition when the effective primary inductance is 100 which is the inductance of the transformer half-primary. T curves marked "combined" show the results of the addition the local feedback developed in the first stage.

Resistive loading on the secondary, by decreasing the amplitu of the output stage whilst at the same time improving the pha relationship, makes the stability margin greater than with t

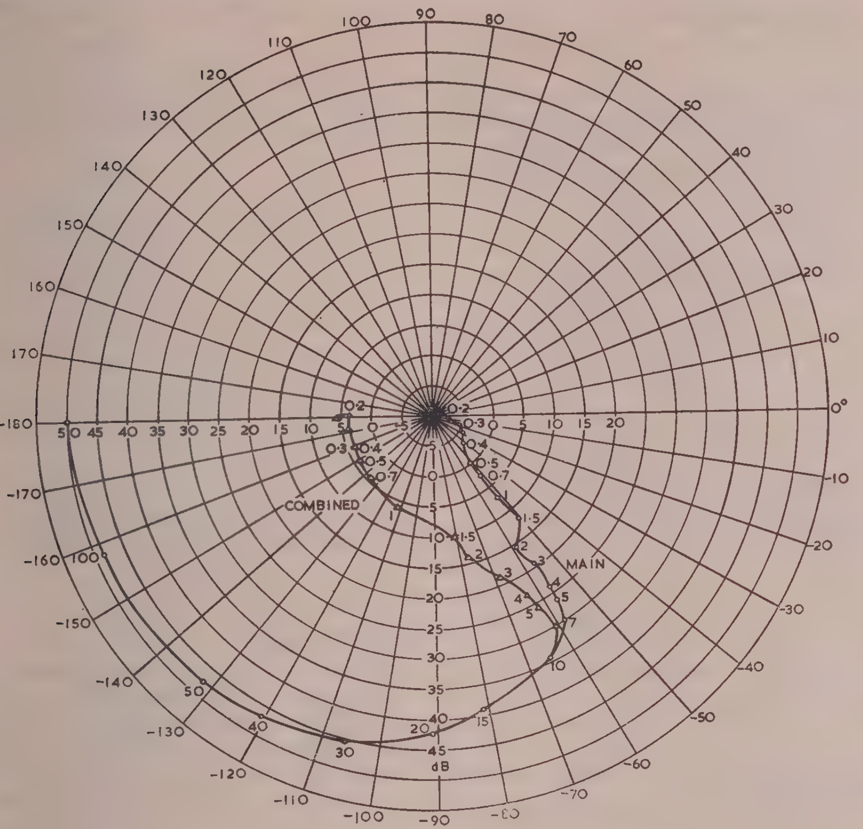


Fig. 5.—Nyquist diagram for low frequencies.

Amplifier termination equivalent to 100 H on each half of transformer primary. Frequency in cycles per second. Results by measurement and calculation.

y inductance alone. Further inductive loading on the rmer secondary may at the most decrease the stability to that shown in Fig. 4, whilst the addition of any able value of capacitance will have negligible effect on the t plot in the critical frequency region.

Grid Current.

n the driver stage is terminated by a resistor having a approximately equal to the minimum grid resistance of put stage, the response of the main feedback loop in the frequency region is unaffected. Whilst this only imates to actual conditions, no instability is shown on s of the waves when the output stage is driven into grid .

Distortion.

n the amplifier is driven at a high frequency without feed- appreciable quantities of harmonics up to the seventh nd in the output, and the harmonic feedback is maintained ficiently high value up to 35 kc/s to reduce these harmonics. circuit is impracticable for this purpose because of the osses that would occur.

Increased Harmonic Feedback.

n the amplifier is operated in class AB_2 the output valves off for a part of each cycle; this sudden change of current o cause ringing, owing to the resonances set up between the winding capacitances and the inductances in the circuit. This tendency to ring is increased as the harmonic feedback is increased.

Although it is desirable to by-pass, in the working band, the local feedback developed across the cathode resistor of the first stage, it was found that, of four circuits designed to satisfy the stability criterion, three caused the output to ring on full drive. Whilst the other was satisfactory in this respect the margin of stability was appreciably less than that of the circuit in use.

However, since the performance of the amplifier is considered satisfactory, with the possible exception of the hum level, the large margin of stability of the circuit in use is at present preferred to the advantages which might arise from the additional 10 dB of harmonic feedback.

(3.7.6) **Feedback Resistors.**

The output-voltage constancy of the amplifier is largely determined by the feedback resistors, which should therefore be free from random variations and have a small temperature coefficient.

The feedback resistance R_{50} is necessarily high because of the large time-constant required in the feedback circuit at low frequencies and the high alternating voltages in the circuit. It has a time-constant of the order of $-0\cdot 3\,\mu$H/ohm and is made from 50 s.w.g. Evanohm enamelled wire having a temperature coefficient of less than 10 parts in 10^{-6} per deg C, which is wound in a single layer on a mica card. The resistor R_2 is made from 40 s.w.g. Evanohm wire of similar temperature coefficient.

(3.8) Performance of Amplifier

Input.—The amplifier develops maximum rated voltage with an input of 0·7 volt.

Output.—At unity power factor the amplifier will develop 1 kVA at 1 kc/s, falling to 700 VA at the extremes of the frequency range 30 c/s–5 kc/s at voltages of 115, 57·5 or 28·75 volts.

If the power factor is less than unity the maximum apparent-power rating is reduced, the values for frequencies of 50 c/s and 1 kc/s being shown in Fig. 6.

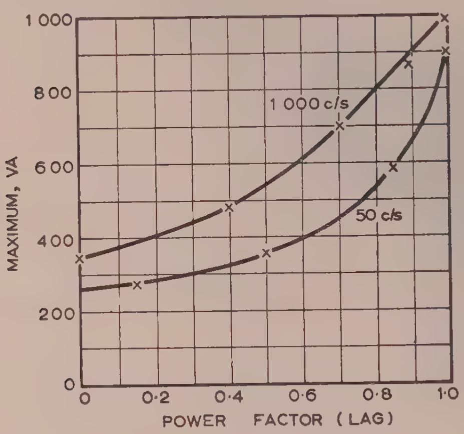


Fig. 6.—Maximum apparent-power rating.

If a single output of greater power is required the amplifier outputs can be connected in series.

Stability.—The amplifier is adequately stable under all conditions of load.

Hum and Noise Level.—The hum and noise output voltage measured on the 115-volt range with the input short-circuited is 0·003 volt.

Output Impedance.—The output impedance, Z_0, on the 115-volt range is given approximately by the expression $Z_0 = 0{\cdot}5 + j\omega \times 10^{-4}$ ohm.

Harmonic Content.—The harmonic distortion was determined at frequencies of 30 c/s, 50 c/s, 1 kc/s and 5 kc/s at various loads and output voltages, and some typical values are given. Table 1 shows the r.m.s. value of the harmonic content, expressed as a percentage of the fundamental when the output voltage is 115 volts and the load is resistive.

Table 1

Power output	Harmonic content			
	at 30 c/s	at 50 c/s	at 1 kc/s	at 5 kc/s
watts	%	%	%	%
0	0·10	0·05	0·00	0·00
300	0·10	0·05	0·05	0·15
500	0·10	0·05	0·05	0·35
700	0·10	0·10	0·10	0·45
900		0·15	0·10	
1 000			0·10	

Gain Constancy.—The change in the amplifier gain does not exceed 0·02% per hour under open-circuit conditions. When delivering maximum power there is an initial fall in gain not exceeding 0·003% per minute.

(4) PERFORMANCE OF OSCILLATOR-AMPLIFIER

(4.1) Harmonic Content

The harmonic distortion was determined at 30 c/s, 50 c 1 kc/s and 5 kc/s at various loads and output voltages, and so typical values are given. Table 2 shows the r.m.s. value of t

Table 2

Power output	Harmonic content			
	at 30 c/s	at 50 c/s	at 1 kc/s	at 5 kc/s
watts	%	%	%	%
0	0·35	0·20	0·15	0·15
300	0·35	0·20	0·15	0·20
500	0·35	0·20	0·15	0·35
700	0·35	0·25	0·15	0·45
900		0·30	0·20	
1 000			0·20	

harmonic content, expressed as a percentage of the fundament when the output voltage is 115 volts and the load is resistive.

(4.2) Voltage Constancy

(4.2.1) Dependence on the Supply Voltage.

The response of the amplifier output when changes of 2 and 6% respectively were made in the supply voltage to bo oscillator and amplifier were recorded when the frequency w set at 60 c/s, the output voltage was 110 volts and the load w 19 ohms non-inductive.

The oscillograph record shown in Fig. 7 was obtained

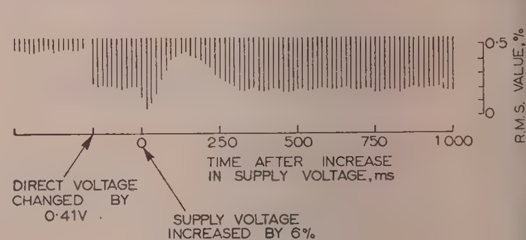


Fig. 7.—Response of output voltage to a 6% supply-voltage chang

applying to the oscillograph a stable d.c. voltage of such magnitude that it "backed-off" the greater part of the sign from the amplifier, leaving only the peaks of the waves on t screen. A reference point, which also served as a calibrati deflection, was made on the record by changing the level of t direct voltage by a known amount approximately one-fifth of second before the change in the supply voltage took place.

The record gives the response of the output when the supp voltage is increased by 6% and shows that the voltage increas by 0·14% during the first 30 millisec, then decreases regular during the next 130 millisec to a minimum, 0·26% below t initial level, and finally increases uniformly to reach the n steady value, 0·03% higher than the original level at the end approximately 400 millisec. The response for the 2% chan follows a similar pattern and the recovery period is of the sam duration. The respective maxima and minima are 0·04% high and 0·08% lower than the initial voltage.

The changes in the output, as indicated by an electrosta voltmeter when the output had steadied after the change supply voltage, were 0·03% for a 6% change and 0·01% for 2% change.

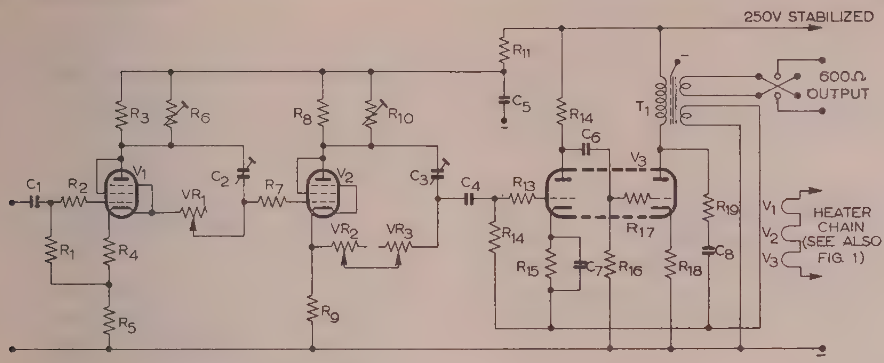


Fig. 8.—Phase-shift unit.

R_1 1 MΩ	R_{11} 2·2 kΩ	VR_1, VR_2 100 kΩ	C_7 100 μF
R_2 1·5 kΩ	R_{12} 1 MΩ	VR_3 200 Ω	C_8 2 200 μμF
R_3 See Note 1*	R_{13} 1·5 kΩ		
R_4 390 Ω	R_{14} 47 kΩ		
R_5 10 kΩ	R_{15} 390 Ω		V_1, V_2 UF42
R_6 See Note 2	R_{16} 1 MΩ	C_1 0·1 μF	V_3 UCC 85
R_7 1·5 kΩ	R_{17} 1·5 kΩ	C_2, C_3 Selected for range	
R_8 See Note 1*	R_{18} 390 Ω	C_4 0·1 μF	
R_9 10 kΩ	R_{19} 2·2 kΩ	C_5 200 μF	T_1 Output 3 : 1
R_{10} See Note 2		C_6 0·5 μF	Feedback 3 : 0·8

* 0·001 in. Evanohm on 0·01 in mica card.
Note 1. Determined for highest frequency range.
Note 2. Determined for other frequency ranges.

Variation with a Constant Supply Voltage.

en the supply voltage is constant to ±0·1% the output as ved by the electrostatic voltmeter is almost entirely free random variations, and variations as great as 0·01% have een observed.

ere is a slow change the value of which is dependent on conditions and does not exceed 0·003% per minute.

Supply-Frequency Component.

supply-frequency component at maximum output does xceed 0·01%.

(5) PHASE-SHIFT UNIT

ttmeters are usually tested at the N.P.L. by the artificial- method, and for this the phase-shift unit is connected en the oscillator and one amplifier, as shown in Fig. 9, at the phase of this output may be varied with respect to ther. A calibrated shift is unnecessary, as the equivalent factor can be obtained with sufficient accuracy from the ard instrument, providing the output of the unit is constant .

circuit employed is shown in Fig. 8. It is similar to that bed in a previous paper[1] and is of the form in which a etwork is connected across the voltages of $-v$ and $+v$ oped in a phase-splitting circuit. It is impracticable to both C and R, so that R only is made variable and two ts, each giving not less than 90°, are used. The reversing in the output stage enables the phase to be varied through her 180°.

output from each network should not change by more 0·5% when the phase is moved through 90°, and in order tisfy this requirement the impedance of the CR circuit d be large compared with that of the valve. The variable ors VR_1 and VR_2 are made as large as the normal range of onents will allow, and the frequency range, 20–5 120 c/s is d into eight octaves with a separate capacitor used for each. connection of a high-impedance network makes the n dependent on stray capacitances; it is found that these be sufficient to cause serious errors above a few hundred cycles per second. Effective compensation for the effects of this capacitance is given over each octave by feeding to the circuit voltages slightly asymmetric in value with respect to earth. The variation of amplitude with phase is reduced to less than 1% over the complete frequency range by changing the anode loads of the valves when the capacitance for the frequency in use is selected. The use of a separate capacitor for each range also gives a more uniform change of phase with resistance.

The change of deflection of a wattmeter at low power factors is directly proportional to the change in angle, so that a fine

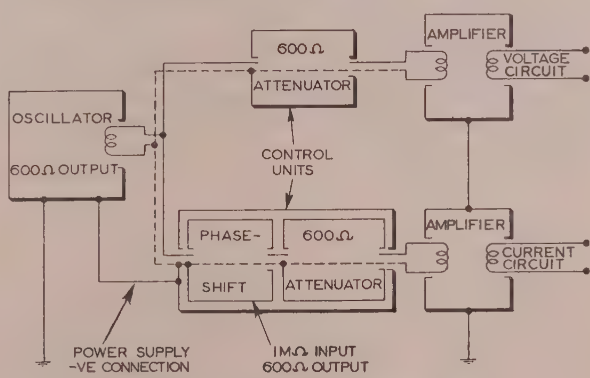


Fig. 9.—Connections for a double supply.

control is therefore required for setting zero power factor. This is provided by VR_3. It is also essential, for the same reason, that both resistances and capacitances in the CR networks should be of good quality and free from small variations that would otherwise cause unsteadiness at this setting.

The two phase-shifting stages are followed by a two-stage amplifier across which 36 dB of negative feedback is developed, and the overall gain of the system is nearly unity.

The valve heaters are connected in series with those of the oscillator and its stabilized power supply, and the h.t. voltage

is also taken from this supply. As in the case of the oscillator, additional filtering is required in the h.t. supply to V_1 and V_2 to reduce the supply-frequency component to negligible proportions.

The distortion introduced by this phase-shift unit is negligible and does not affect the total distortion in the amplifier output. The voltage constancy is at least equal to that of the remainder of the equipment.

This phase-shifting system, unlike that dependent on either a phase-shifting capacitor, a phase-shifting transformer or a sine-cosine potentiometer, does not depend for its amplitude constancy on a precise mechanical system, and in this respect is simpler and easier to construct. It is unsuitable, however, when it is desired to introduce a specified phase-change at a number of frequencies.

(6) EARTHING ARRANGEMENTS

Care has been taken in the design of the individual units to ensure that the supply-frequency components in each are small. The extent to which this has been achieved is indicated in the performance figures given, and it is essential that they are not increased by the system of interconnection.

Screening against electromagnetic radiation is provided by using concentric cable for interconnecting, and to ensure that it is effective the sheath is earthed at one point only. Precautions are also taken to avoid currents of supply frequency flowing through it to earth. The oscillator output and amplifier inputs are therefore isolated from the chassis and, in the double-output arrangement shown in Fig. 9, the earth connection to the sheath is made at the phase-shift unit by the negative connection of the common power supply.

(7) CONCLUSIONS

Electronic supplies are now being used extensively at the N.P.L. for the calibration of instruments, and five amplifiers, three oscillators and two phase-shift units have been constructed to the designs given in the paper.

The oscillator-amplifier set described has been in operation for almost a year, and during this period the performance figures have been maintained. Experience has shown, therefore, that high-power electronic supplies for instrument calibration are both practicable and reliable.

(8) ACKNOWLEDGMENTS

The work described has been carried out as part of the resear[illegible] programme of the National Physical Laboratory, and this pap[illegible] is published by permission of the Director of the Laboratory.

The authors desire to acknowledge the comments and su[illegible]gestions received from Mr. P. H. Taylor of Messrs. W. Bry[illegible] Savage Ltd., and Mr. W. Wilson of Control Mechanisms an[illegible] Electronics Division, National Physical Laboratory. Mr. R. [illegible] Chalmers of Control Mechanisms and Electronics Divisio[illegible] National Physical Laboratory, is responsible for the design of t[illegible] output transformer.

(9) REFERENCES

(1) Arnold, A. H. M.: "Alternating-Current Testing Equ[illegible]ment," *Proceedings I.E.E.*, Paper No. 1532 M, July, 19[illegible] (**101**, Part II, p. 121).

(2) Cooper, W. H. B., and Seymour, R. A.: "Temperatu[illegible] Dependent Resistors. Use as Electric Circuit Elements[illegible] *Wireless Engineer*, 1947, **24**, p. 298.

(3) Mezger, Lt. C. R.: T.M.B. Report 502.

(4) Benson, F. A.: "Voltage Stabilizers," *Electronic En[illegible]neering*, London, 1950, p. 67.

(5) Reich, H. J.: "Theory and Application of Electron Tube[illegible] (McGraw-Hill, New York, 1944), p. 593.

(6) Lord, H. W.: "The Design of Broad-Band Transformers f[illegible] Linear Electronic Circuits," *Transactions of the Americ[illegible] I.E.E.*, 1950, **69**, Part II, p. 1005.

(7) McLean, T.: "An Analysis of Distortion in Class B Au[illegible] Amplifiers," *Proceedings of the Institute of Ra[illegible] Engineers*, 1936, **24**, p. 487.

(8) Barton, L. E.: "Recent Developments of the Class [illegible] Audio and Radio Frequency Amplifiers," *ibid.*, p. 985.

(9) Rockwell, R. J., and Platts, G. F.: "Automatic Co[illegible]pensation for Class B Bias and Plate Voltage Regulatio[illegible] *ibid.*, p. 553.

(10) Valley, G. E., and Wallman, H.: "Vacuum Tube Am[illegible]fiers" (McGraw-Hill, New York, 1948), Chapter 9.

(11) Bode, H. W.: "Network Analysis and Feedback Ampli[illegible] Design" (D. Van Nostrand, New York, 1945).

(12) Duerdoth, W. J.: "Some Considerations in the Design [illegible] Negative Feedback Amplifiers," *Proceedings I.E.[illegible]* Paper No. 851 R, October, 1949 (**97**, Part III, p. 138).

(13) Learned, V. L.: "Corrective Networks for Feedb[illegible] Circuits," *Proceedings of the Institute of Radio Engine[illegible]* 1944, **32**, p. 403.

BROADBAND TRANSISTOR FEEDBACK AMPLIFIERS

By J. ALMOND, M.Sc., and A. R. BOOTHROYD, Ph.D., Graduates.

(*The paper was first received 6th April and in revised form 22nd July, 1955.*)

SUMMARY

possibilities of negative-feedback amplifiers involving three ed common-emitter junction-transistor stages are investigated. approximate expressions are given for the power gain, input ance and output impedance of series, shunt and compound ck amplifiers, for the case of large applied negative feedback. hown that with large negative feedback, dynamic stability conions may be analysed with little error by considering the forward eturn paths of the feedback loop as separable, in spite of the ction effects between transistors. Owing to the restricted frey range of application of the transistors employed, the maximum ve feedback possible without instability depends almost exclu- on the frequency characteristics of the current gain, α, of the stor; it is found that the α-characteristics may be represented adequate accuracy by a minimum-phase (RC) approximation, leads to simple analysis both of stability conditions and of le phase and amplitude margins. Examples are given of feed- amplifiers having overall operating power gain and applied ve feedback of approximately 33 dB and 30 dB respectively, in dynamic stability is secured by using phase-advancing networks return path of the feedback loop.

LIST OF SYMBOLS

= Transistor current-amplification parameter $= -\partial I_c/\partial I_e$.
= Transistor current gain in the common-emitter connection, ideally with zero collector load impedance. $b \simeq \alpha/(1-\alpha)$.
= Amplifier current gain.
= Amplifier voltage gain.
= Feedback amplifier current gain with feedback removed.
= Feedback amplifier voltage gain with feedback removed.
= Amplifier power gain.
= Amplifier operating power gain.

= Transistor low-frequency parameters.

= Generator source resistance.
= Amplifier load resistance.
= Series feedback resistance.
= Shunt feedback resistance.
= Amplifier input resistance.
= Amplifier output resistance.
= Return ratio.
= Return difference.
= Amplifier mesh determinant.
= Mesh determinant with one transistor inoperative.
= Low-frequency value of α.
= Frequency at which $|\alpha| = 0{\cdot}707\alpha_0$.
= Low-frequency value of b.
= Frequency at which $|b| = 0{\cdot}707b_0$.
= Fraction of output voltage fed back in the positive sense.
= Output voltage.
= Generator voltage.
= Input current.
= Output current.

en contributions on papers published without being read at meetings are for consideration with a view to publication.
Almond is with the Defence Research Board of Canada, Ottawa.
oothroyd is in the Electrical Engineering Department, City and Guilds , University of London.

(1) INTRODUCTION

The present state of transistor development is such that considerable gain stabilization is necessary in any high-grade amplifier application. This is particularly so in the case of wide-band amplifiers of the repeater type, where with established valve techniques it is customary to apply large negative feedback. Little published work has yet appeared on transistor feedback amplifiers, although Shea[1] has discussed possible circuit arrangements and has indicated methods of investigating the dynamic stability of such amplifiers. The paper goes into certain basic design aspects of transistor amplifiers in which large negative feedback, e.g. 30 dB, is applied.

The design of transistor feedback amplifiers is in some respects more involved than with thermionic valves owing to the physical nature of the transistor itself. Even at low frequencies the transistor is represented by a more complicated equivalent circuit than that of a valve, and therefore circuit analysis is also more complicated; in particular, the input impedance of a transistor amplifier stage is in general low, and is also considerably affected by conditions in its output circuit and by subsequent stages. Thus in a feedback amplifier where the amount of feedback applied is influenced by the input circuit impedance, exact analysis is complicated by the impossibility of separating the forward and return paths, and results in very unwieldy expressions for circuit performance. It is shown in the paper, however, that useful yet fairly simple expressions may be derived if certain approximations are made. These approximations depend on the different orders of magnitude of the various transistor parameters and require that the applied negative feedback be large.

An important consideration in feedback amplifier design is dynamic stability. With valves, stability is governed by linear minimum-phase-shift networks, e.g. by parasitic capacitance at high frequencies. But transistors introduce an additional effect, namely, the conduction process through the semi-conductor from emitter to collector which, owing to transit-time dispersion of current carriers, is responsible for a frequency-dependent transistor current gain α. Now the functional form of α in terms of frequency is very complicated, even for a transistor of simple physical shape, and may not be represented exactly by an equivalent circuit. Thomas,[2] however, has shown that over a limited frequency range a reasonable approximation to the representation of α is given in terms of a simple RC minimum-phase-shift network—i.e. a single time-constant. The frequency-dependence of α affects the stability of a feedback amplifier at relatively high frequencies; in fact, so low is the frequency range of effectively constant current-gain in many transistor amplifier stages that α is the dominant factor controlling high-frequency stability, effects due to parasitic capacitance being negligible. The paper aims mainly to show that, in practice, if the feedback amplifier consists basically of cascaded common-emitter stages, Thomas's approximation for the frequency-dependence of α is sufficiently good for reliable design to be carried out; also that high-frequency stability may be secured with large amounts of applied negative feedback by compensating the phase characteristic of α by simple phase-advance networks.

In the following discussion, common-emitter junction-transistor amplifier stages are assumed throughout. Junction transistors are

usually preferable to the point-contact type for amplifier applications owing to their superior noise figure and the far greater linearity of their static characteristics. Moreover, the junction transistor can be used to give useful gain in three different modes of connection. On account of the inter-stage impedance-matching necessary to realize a power gain with the common-base connection, base input arrangements appear to be the most appropriate for wide-band amplifier applications, since direct or *RC* coupling between stages may then be used. Of these two possibilities, the common-emitter connection gives by far the greater power gain, and only this type of amplifier stage is considered.

The current gain of a common-emitter stage with a small load impedance is approximately $\alpha/(1-\alpha)$, where α is the current-amplification parameter of the transistor, defined in the common-base configuration as $-\partial I_c/\partial I_e$. For convenience this current gain is given the symbol b. Thus if common-emitter stages are cascaded with direct or *RC* coupling the power gain per stage is approximately b^2. For typical transistors α is approximately 0·95 giving a power gain of about 26 dB per stage. However, quite small changes in α result in relatively large changes in power gain: for example, if α falls to 0·90, the power gain per stage is reduced to about 19 dB. If such changes in α are to be expected over a period of time or a range of transistors, it would mean that, in order to achieve reasonable gain-stability with a single-stage amplifier, almost all the available gain of the amplifier would need to be utilized for the application of negative feedback. For most practical purposes, therefore, feedback over more than one stage is necessary, and as the common-emitter stage gives a phase reversal, the most suitable number of stages is three if a phase-inverting circuit is not utilized. In the following Sections the application of large amounts of negative feedback to a basic 3-stage *RC*-coupled common-emitter amplifier is discussed.

(2) THREE-STAGE COMMON-EMITTER AMPLIFIER WITHOUT FEEDBACK

The properties of a three-stage common-emitter amplifier are first to be considered without feedback applied. No attempt is made to provide matching between stages, so that the frequency range of the amplifier is controlled only by the transistors. The only interstage networks are coupling capacitors and resistors, to supply and isolate the d.c. operating potentials. Fig. 1 shows,

Fig. 1.—Basic 3-stage amplifier arrangement.

schematically, the arrangement of the 3-stage amplifier. This is an idealized circuit showing no coupling capacitors or resistors and is the form the circuit will be assumed to take in the middle frequency range. The exact analysis of this circuit is rather involved, but by making certain simplifying approximations useful relations can easily be developed on the assumption that all transistors have the same small signal parameters.

The first two transistors work into a low load impedance (i.e. the input impedance of the next stage). Therefore, the current gain of the first two transistors is approximately equal to the short-circuit current gain, b, of the common-emitter stage [i.e. $b=\alpha/(1-\alpha)$], and the current gain for the last stage is approximately equal to $\alpha/(1-\alpha+R_L/r_c)$. The current gain for the complete amplifier is, therefore, given by

$$K_i=\frac{i_o}{i_i}=\left(\frac{\alpha}{1-\alpha}\right)^2\frac{\alpha}{1-\alpha+R_L/r_c}\simeq\left(\frac{\alpha}{1-\alpha}\right)^3 \quad . \quad (1)$$

The approximation in this equation is valid if R_L is small com-pared with $r_c(1-\alpha)$. The input impedance R_i of the amplifi- is approximately equal to the short-circuit input impedance o the first stage, i.e.

$$R_i\simeq r_e+r_b+r_e\alpha/(1-\alpha) \quad . \quad . \quad . \quad . \quad (2$$

Thus the output voltage is

$$v_o\simeq-\left(\frac{\alpha}{1-\alpha}\right)^3 i_iR_L$$

if R_L is assumed small compared with $r_c(1-\alpha)$.

Putting $i_i=\dfrac{v_g}{R_i+R_g}$ in this equation, the voltage gain is

$$K_v=v_o/v_g\simeq-\left(\frac{\alpha}{1-\alpha}\right)^3\frac{R_L}{R_g+r_e+r_b+r_e\alpha/(1-\alpha)} \quad . \quad (3$$

If the power gain G is defined as the power delivered to the loa divided by the power delivered to the input of the transistor

$$G=\left(\frac{i_o}{i_i}\right)^2\frac{R_L}{R_i}\simeq\left(\frac{\alpha}{1-\alpha}\right)^6\frac{R_L}{r_e+r_b+r_e\alpha/(1-\alpha)} \quad . \quad (4$$

The operating power gain G_o is therefore given by

$$G_o=K_v^2\,4\frac{R_g}{R_L} \quad . \quad . \quad . \quad . \quad . \quad . \quad (5$$

The approximate current gain of the amplifier given by eqn. (1 is of the order of 75 dB for typical transistors. If the value o the load impedance R_L is made equal to the input impedanc this gives power and voltage gains of 75 dB for the three stage Now if 30 dB of this gain is utilized for negative feedback, t stabilize the gain, the amplifier still gives a power gain of 45 dI assuming that the input impedance of the amplifier is not change by the application of feedback.

(3) PROPERTIES OF SERIES, SHUNT AND COMPOUN FEEDBACK AMPLIFIERS

Series and shunt feedback are shown simultaneously applie to a 3-stage amplifier in Fig. 2, the elements R_1 and R_2 bein

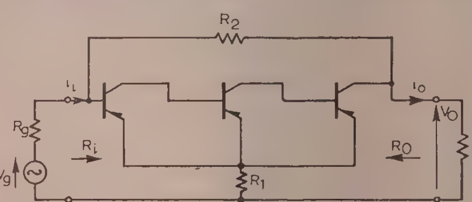


Fig. 2.—Application of series, shunt and compound feedback to t basic 3-stage amplifier.

responsible for the two effects respectively. The coupling a by-pass capacitors and biasing resistors are omitted, as before.

The circuit mesh equations for this amplifier are readily writt down in terms of the equivalent T-network of the transisto Exact solution of these equations for amplifier gain, inp impedance and output impedance results in extremely comp cated expressions, however, in which many terms are in fa quite negligible. Useful approximate expressions, involvi small errors only, result if certain assumptions are mac Suppose that all transistors have the same parameter valu and that r_b, r_e, R_1 and R_L are small compared with (r_c-r_m also that R_2 is comparable in magnitude with r_c and r_m a that R_g is much smaller than ($R_2+R_L+R_g$) and (r_c-r Although this leads to considerable simplification, the expr

concerned are still rather involved; the approximate ...sions for amplifier properties with and without feedback ...d are set out in Table 1. Note that the shunt feedback ... has the series-feedback element R_1 short-circuited, while ... series feedback circuit R_1 is the feedback element and R_2 ...n-circuited. The compound feedback circuit is in the ...of Fig. 2. All quantities given in Table 1 have already ...defined with the exception of $T_{2,3}$. This quantity is the return ratio of the amplifier for the second- or third-stage transistors and will be discussed later. The expressions in Table 1 are general to the extent that there is no restriction on the amount of negative feedback applied to the amplifier. Before discussing the amplifier properties, however, it is advantageous to make further approximations that may be applied when the amount of negative feedback is large, which is the main concern of the paper. As discussed later in this Section, if there

Table 1

Approximate Expressions for Amplifier Properties

	Without feedback	Shunt feedback	Series feedback	Compound feedback
R_i	$R_3 = r_e(1+b) + r_b$	$\dfrac{R_3}{1 + \dfrac{R_L}{R_2 + R_L} b^3}$	$(R_1 + R_3)\left(1 + \dfrac{R_1 b^3}{R_1 + R_3}\right)$	$\dfrac{(R_1 + R_3)\left(1 + \dfrac{R_1 b^3}{R_1 + R_3}\right)}{1 + \dfrac{R_L}{R_2 + R_L} b^3}$
R_o	$r_c(1-\alpha)$	$\dfrac{r_c(1-\alpha)}{1 + \dfrac{r_c(1-\alpha)R_g}{(R_2 + R_g)(R_g + R_3)} b^3}$	$r_c(1-\alpha)\left(\dfrac{R_1 b^3}{R_1 + R_g + R_3}\right)$	$\dfrac{r_c(1-\alpha)\left(1 + \dfrac{R_1 b^3}{R_1 + R_g + R_3}\right)}{1 + \dfrac{r_c(1-\alpha)(R_1 + R_g)b^3}{(R_2 + R_g)(R_1 + R_g + R_3)}}$
K_v	$-b^3 \dfrac{R_L}{R_g + R_3}$	$\dfrac{-b^3 R_2 R_L}{(R_2 + R_L)\left(R_g + R_3 + \dfrac{R_g R_L}{R_2 + R_L} b^3\right)}$	$\dfrac{-b^3 R_L}{R_1 + R_g + R_3 + R_1 b^3}$	$\dfrac{-R_2 R_L b^3}{(R_2 + R_L)\left[R_1 + R_3 + R_g + \left(R_1 + \dfrac{R_g R_L}{R_2 + R_L}\right) b^3\right]}$
K_i	b^3	$\dfrac{b^3 R_2}{(R_L + R_2)\left(1 + \dfrac{R_L}{R_2 + R_L} b^3\right)}$	b^3	$\dfrac{R_2 b^3}{(R_2 + R_L)\left(1 + \dfrac{R_L b^3}{R_2 + R_L}\right)}$
G	$b^6 \dfrac{R_L}{R_3}$	$\dfrac{b^6 R_2^2 R_L}{(R_L + R_2)^2\left(1 + \dfrac{R_L}{R_2 + R_L} b^3\right) R_3}$	$\dfrac{b^6 R_L}{R_1 + R_3 + R_1 b^3}$	$\dfrac{b^6 R_2^2 R_L}{(R_2 + R_L)^2\left(1 + \dfrac{R_L b^3}{R_2 + R_L}\right)(R_1 + R_3 + R_1 b^3)}$
$T_{2,3}$		$\dfrac{R_g R_L b^3}{(R_2 + R_L)(R_g + R_3)}$	$\dfrac{R_1 b^3}{R_1 + R_g + R_3}$	$\dfrac{b^3\left(R_1 + \dfrac{R_g R_L}{R_2 + R_L}\right)}{R_g + R_1 + R_3}$

$b = \alpha/(1-\alpha)$
$R_3 = r_e(1+b) + r_b$

Table 2

Approximate Expressions for Amplifier Properties with Large Negative Feedback Applied

	Shunt feedback	Series feedback	Compound feedback
	$\dfrac{R_3(R_2 + R_L)}{R_2 b^3}$	$R_1 b^3$	$\dfrac{R_1(R_2 + R_L)}{R_L}$
R_o	$\dfrac{(R_2 + R_g)(R_3 + R_g)}{R_g b^3}$	$\dfrac{r_c(1-\alpha)R_1 b^3}{R_1 + R_g + R_3}$	$\dfrac{R_1(R_2 + R_g)}{R_1 + R_g}$
K_v	R_2/R_g	R_2/R_1	$\dfrac{R_2 R_L}{R_1(R_2 + R_L) + R_g R_2}$
K_i	R_2/R_L	b^3	R_2/R_L
G	$b^3 \dfrac{R_2^2}{(R_2 + R_L)R_3}$	$b^3 \dfrac{R_L}{R_1}$	$\dfrac{R_2^2}{(R_2 + R_L)R_1}$

$b = \alpha/(1-\alpha)$
$R_3 = r_e(1+b) + r_b$

is a large amount of negative feedback, the return ratio T may be written with negligible error

$$T \simeq -\beta K_{v0}$$

where K_{v0} is the amplifier voltage gain without feedback and β is the fraction of the output voltage fed back to the input. If the return ratio of the amplifier, due to the feedback elements R_1 and R_2, is large compared with unity (i.e. $\geqslant 30$), further simplification of the expressions of Table 1 is justified, giving the expressions listed in Table 2 for the amplifier properties.

It can be seen from Table 2 that the voltage gain K_v is in all three cases virtually independent of small changes of transistor parameters. The input and output impedances are now dependent on the factor b^3 in both series and shunt feedback. Since this factor is most sensitive to changes in α, this may be a serious drawback in practical circuits. The power gain for series and shunt feedback is still proportional to b^3; this is some improvement, however, since without feedback the power gain is proportional to b^6. It can be seen from the last column in the Table that with large amounts of negative feedback applied simultaneously by both series and shunt methods all properties of the amplifier are almost independent of changes in transistor parameters.

The quantities in Table 2 have been calculated for a typical set of transistor parameters and are set out in Table 3.

Table 3

CALCULATIONS FOR 3-STAGE AMPLIFIER

$\alpha = 0{\cdot}94$; $r_c = 750$ kilohms; $R_2 = 70$ kilohms; $R_1 = 7$ ohms; $R_g = 100$ ohms; $R_L = 5$ kilohms; $R_3 = 700$ ohms.

Parameter	Without feedback	Shunt feedback	Series feedback	Compound feedback
R_i (ohms)	700	2·75	26 600	105
R_o (ohms)	45 000	147	$1{\cdot}48 \times 10^6$	4 640
K_v	$2{\cdot}29 \times 10^4$	700	715	341
K_i	$3{\cdot}82 \times 10^3$	14	$3{\cdot}8 \times 10^3$	14
G	$1{\cdot}04 \times 10^8$	$3{\cdot}55 \times 10^5$	$2{\cdot}72 \times 10^6$	$9{\cdot}35 \times 10^3$
$T_{2,3}$		31·8	32·8	64·5

In the second column the properties of a typical 3-stage amplifier without feedback are given. The feedback resistors, R_1 and R_2, are chosen to give approximately equal amounts of series or shunt feedback. The last column refers to an amplifier in which both types of feedback are applied simultaneously.

The symbol $T_{2,3}$ in Tables 1 and 3 stands for the return ratio for the second or third transistor in the amplifier circuit. The return ratio T and the return difference F can be found for any transistor in the circuit from considerations analogous to those discussed by Bode[3] for valve amplifiers. T and F are defined by eqns. (6) and (7):

$$F = \Delta/\Delta_0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

$$T = F - 1 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

where Δ is the determinant formed by the circuit mesh equations and Δ_0 is the determinant resulting from Δ when the transistor under consideration is made inoperative. The manner in which this is done for the calculation of Δ_0 is discussed by Shea,[1] and requires a little more explanation. In the derivation of eqns. (6) and (7), the current which is fed back to the input terminals of the amplifier (in this case to the base of the first transistor), due to an initial current entering at that terminal, is found. This is done by assuming that the voltage appearing in the output mesh of the first transistor stage, due to an initial current in the input mesh, acts as an externally applied voltage. Thus, if the

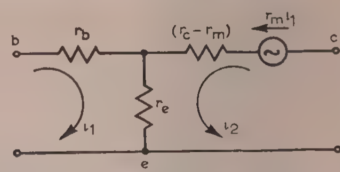


Fig. 3.—Common-emitter equivalent-T circuit of the transistor.

equivalent circuit for the common-emitter transistor shown Fig. 3 is used, Δ_0 is obtained by simply making $i_1 r_m$ equal zero, the self-impedance element $(r_c - r_m)$ of the second me remaining unchanged.

If the return ratios, T, for the transistors in the circuit co sidered here are calculated, there appears a small differen between the return ratio for the first transistor and the retu ratios for the second and third transistors. This is due to t fact that the generator impedance has been assumed to be the same order of magnitude as the input impedance of the fi stage, so that the effect of the internal negative-feedback resistan r_e is appreciable. The source impedance for the second a third stages is the output impedance of the preceding sta which is much larger than the input impedance; this rende the local feedback due to r_e negligible in these stages. If, ho ever, the feedback around the complete amplifier is relativ large (e.g. $T \geqslant 30$), the local feedback due to r_e can be neglect in all three stages.

Although the return ratio T and the return difference F ha been used in the above discussion for the circuits considered he the return ratio $T_{2,3}$, obtained from eqn. (7), together with t approximations discussed at the beginning of this Section, reduc effectively to the loop gain $(-K_{v0}\beta)$, where K_{v0} is the voltage ga without feedback and β is the fraction of output voltage fed bac obtained by using the simple feedback theory in which separa forward and return paths are assumed. The amplifier loop ga can therefore be considered when investigating the dynam stability of the circuit.

(4) CONSIDERATIONS OF DYNAMIC STABILITY

(4.1) High-Frequency Instability due to Transistor α-Characteris

Referring to the circuit shown in Fig. 2, it can be seen th when the feedback element is purely resistive the loop ga $(-K_{v0}\beta)$ is proportional to the term $[\alpha/(1-\alpha)]^3$, which effectively the current gain for the three stages without feedbac

The variation of the current gain α with frequency for t junction transistor has been found[2] to be given fairly accurat at frequencies below f_α by

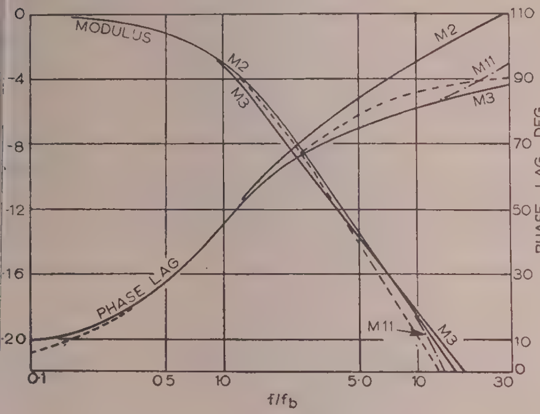
$$\alpha = \frac{\alpha_0}{1 + jf/f_\alpha} \quad . \quad . \quad . \quad . \quad . \quad . \quad .$$

where f_α is the frequency at which α is reduced in magnitude 3 dB compared with its low-frequency value α_0. Substituting f α from eqn. (8), the short-circuit current-gain expression for t common-emitter stage becomes

$$b = \frac{\alpha}{1-\alpha} = \frac{\alpha_0}{1-\alpha_0}\left[\frac{1}{1 + j\dfrac{f}{f_\alpha(1-\alpha_0)}}\right] = \frac{b_0}{1 + jf/f_b} \quad .$$

where b_0 has been written for $\alpha_0/(1-\alpha_0)$ and f_b for $f_\alpha(1 - \alpha$ From this equation it can be seen that the cut-off frequen of the current gain b for the common-emitter stage δ is (1 — c times the α cut-off frequency f_α. For example, with $\alpha_0 = 0{\cdot}$ and $f_\alpha = 500$ kc/s, $f_b = 26{\cdot}4$ kc/s. The above effect of t frequency variation of α on b, was first pointed out by Thoma

$(1 - \alpha_0)$ is very much smaller than unity (e.g. approximately only the variation of b with frequency below f_α is of interest e present application. The variation of b with frequency iree junction transistors is shown in Fig. 4 together with a



4.—Transistor current-gain characteristics in the common-emitter connection (modulus and phase of b).

---- Simple minimum-phase approximations.
—— Actual measurements.

e calculated from eqn. (9). The current gain b is normalized respect to the low-frequency value b_0, and the frequency is ed in terms of f_b. From this curve it can be seen that the ulus curves agree with the calculated curve to within 2 dB the complete frequency interval shown, and for frequencies w $7f_b$ to within 0·5 dB. The phase curves for the transistors and M3 agree with the calculated curves to within 5 degrees the complete frequency range, but there is a serious epancy for transistor M2 at frequencies above $10f_b$. reason for this discrepancy for M2 is that α_0 for this sistor is very low (0·89), so that $(1 - \alpha_0)$ is approximately . The cut-off frequency f_b is thus $0{\cdot}11 f_d$, so that $10f_b$ is er than f_α, and eqn. (8) no longer gives a reasonable approxi- on. The low-frequency values of α for transistors M11 and are 0·95 and 0·96 respectively, so that $10f_b$ is still below f_α these transistors. Since the low-frequency value of α for t transistors is around 0·95, or higher, eqn. (9) would seem a onable approximation for the variation of b, both in phase in modulus.

aking the above assumptions regarding the frequency varia- of α, it is easily seen how much feedback can be applied to stage common-emitter amplifier before the circuit becomes able, when the feedback fraction β is real and constant. modulus of the simple minimum-phase expression given for current gain b by eqn. (9) is 6 dB down from its low-frequency e when the phase change is 60 deg. Three such stages efore give 180° phase change when the modulus of b is ced by 18 dB. Thus if the only frequency-dependent term ne forward gain of the amplifier is $[\alpha/(1 - \alpha)]^3$ the maxi- n possible loop gain, with feedback applied through purely tive elements, is 18 dB. If feedback is defined as referring he quantity analogous to the return difference F—i.e. $K_{v0}B$)—then the feedback under the above conditions is ed to slightly over 19 dB, as with valve amplifiers.

4.2) Control of Stability using Phase-Advance Networks

the amount of feedback were increased beyond the 19 dB mentioned in Section 4.1, the amplifier would become able. For most high-grade applications 19 dB of negative back would be insufficient adequately to control the gain stability of the amplifier. Modification of the circuit, so that more feedback can be applied to the amplifier without instability, could be carried out in the forward or return paths of the amplifier. Modification of the forward path of the amplifier would, however, result in a decrease of the gain of the amplifier in the useful band, or a decrease of useful bandwidth. Since both the useful band and the low-frequency gain of transistor amplifiers are already relatively low, any decrease in these quantities is to be avoided if possible. The amount of feedback can be increased, without affecting the forward gain of the amplifier, by using phase-advancing networks in the feedback circuit.

It was pointed out in Section 4.1 that if the feedback is applied through purely resistive elements the loop gain is proportional to b^3. Assuming that over the range of interest the frequency variation of b is accurately given by eqn. (9), then the high-frequency loop gain will be controlled by the factor

$$\left(\frac{1}{1 + j\dfrac{f}{f_b}}\right)^3$$

The modulus of this factor approaches an asymptote of 18 dB per octave (i.e. 6 dB per octave per stage). It is pointed out in Reference 3 that if the network is of a minimum-phase-shift type, the phase shift is roughly proportional to the rate at which the modulus of the loop gain is reduced. To limit the phase shift associated with the loop gain in the cut-off region, so that more feedback can be applied, the rate at which the loop gain falls off must be decreased. Since the short-circuit current gain of the common-emitter stage can be fairly closely approximated by a minimum-phase-shift network, this reasoning applies here. The simplest method of decreasing the rate at which the loop gain $(-K_{v0}\beta)$ falls off with frequency, without loss of bandwidth or low-frequency gain, is to make the feedback fraction β increase in magnitude with frequency. Before individual circuits are considered, it is useful to find some indication of a theoretical limit that can be achieved with the above method of phase compensation.

For simplicity, only the shunt feedback case is discussed in detail, but the same approach applies for the series and compound feedback amplifiers. If it is assumed that the purely resistive type of feedback gives satisfactory characteristics at low frequencies, any circuit used must reduce at low frequencies to a simple resistance between the collector of the last transistor and the base of the first transistor. The feedback part of this circuit is shown in Fig. 5.

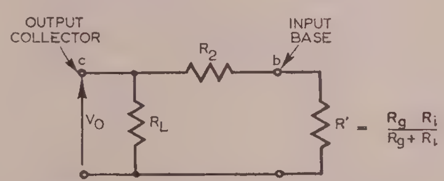


Fig. 5.—Basic shunt feedback network.

It is of interest to find the phase advance that can be achieved with this network, to counteract the phase-retarding effect of the short-circuit current gain. This may be determined by the application of a phase area theorem given by Bode[3] who states (p. 287) "that the total area under the imaginary component [of a minimum-phase-network function] plotted on a logarithmic frequency scale, depends only upon [being equal to $\pi/2$ times] the difference between the values assumed by the real component at zero and infinite frequency." The transfer impedance Z_T of the network under consideration is governed by this theorem, and if

the natural logarithm of the transfer impedance is considered the area concerned will be the area under the phase characteristic. Moreover, if

$$|Z_T|_{\omega=\infty} > |Z_T|_{\omega=0}$$

the area is positive so that the phase curve is of phase-advance type and the available phase-advance area may, with suitable network design, be disposed as desired in order to increase the frequency at which the phase of the loop gain $(-K_{v0}\beta)$ is 180°. The transfer impedance at zero frequency follows from Fig. 5:

$$Z_{T0} = \frac{R_L R'}{R_L + R' + R_2} \quad . \quad . \quad . \quad . \quad (10)$$

At infinite frequency suppose R_2 to be effectively short-circuited so that the maximum phase-advance area is secured; then

$$Z_{T\infty} = \frac{R_L R'}{R_L + R'} \quad . \quad . \quad . \quad . \quad . \quad (11)$$

The circuit in Fig. 6(*a*) has these properties. The difference between Z_{T0} and $Z_{T\infty}$ depends upon R_2, which in turn depends

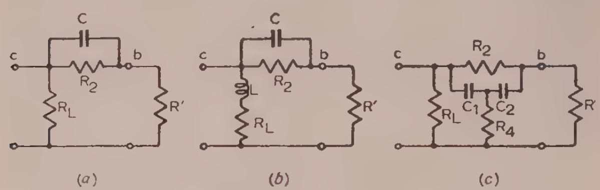


Fig. 6.—Shunt feedback networks incorporating phase advance.

on the amount of feedback required at low frequencies for a given forward gain. It should be pointed out that the smaller the value of R_2, the less the phase area available for compensation; on the other hand, a smaller R_2 implies larger values of feedback, requiring more phase area to make the amplifier stable. For given values of R_L, R_2 and R_1 the amount of phase area available for compensation can be calculated from eqns. (10) and (11). To illustrate this phase-area principle, the phase angle associated with the current gain of a 3-stage amplifier is plotted on Fig. 7. The continuous curve A is the ideal one

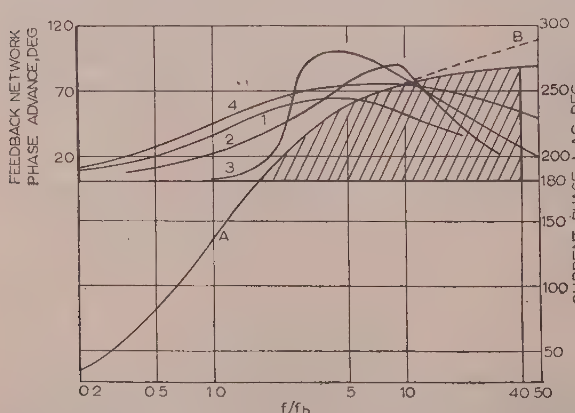


Fig. 7.—Application of phase advance for securing dynamic stability.

obtained by using eqn. (9) and the broken curve B is a more realistic one obtained from measured results. The shaded area on the graph, above the 180° line, corresponds roughly to the phase area available for compensation for values of R_1, R_L and R_2 of 100 ohms 5 kilohms and 70 kilohms respectively. (This value of R_2 would correspond to about 30 dB of feedback with transistors of the type employed.) If the compensating phase area could be arranged just to cancel the shaded portion the phase angle of the loop gain $(-K_{v0}\beta)$ would not cross the 180° line until the frequency had reached $40f_b$. At $40f_b$ the forward gain K_{v0} would be reduced by approximately 32 dB per stage, or 96 dB altogether. The transmission through the loop would be increased by approximately 23 dB at this frequency as compared with zero frequency. The total loop gain would therefore be reduced by 73 dB compared with its low-frequency value when the phase angle had reached 180°. Thus theoretically, if it were possible to produce exactly the phase area shown in Fig. 7 with the β circuit, it would be possible to apply up to 73 dB of feedback.

The above calculations are only for an idealized case, but they serve to show the theoretical limit in the amount of feedback that could be applied in this way, and they also give a measure of the suitability of any network that may be used in the feedback circuit.

(4.3) Phase and Amplitude Margins

It is useful at this stage to consider the modulus and phase margins that would need to be used in a practical amplifier circuit of this type. Clearly the amplifier must be absolutely stable—not conditionally stable—since any decrease in forward gain may make a conditionally stable amplifier unstable. If an upper limit is prescribed for the values of α of the transistors and the amplifier is designed to be stable with this value of α, then only a small modulus margin will be required. The phase margin, however, presents a more difficult problem. In the above calculations the minimum-phase approximation for α has been used, so that adequate allowance must be made for this approximation in the phase margin. From the curves shown in Fig. 4 for the variation of b with frequency, the largest deviation from the calculated curves is approximately 10° up to a frequency of $10f_b$. This deviation is observed for a transistor of low α. If an amplifier is to be designed to give the same gain, and remain stable, with such a transistor in the circuit, or if one of the transistors in the circuit should so deteriorate, then adequate allowance should be made for this extra phase shift. Another possible cause of extra phase shift is the collector capacitance C_c which may be considered as effectively shunting the collector resistance r_c. When R_L is much smaller than $r_c(1-\alpha)$ this effect is negligible. If, however, the last stage has a load impedance of 5 kilohms, as was used in the sample calculations in Section 4.2, small effects will be produced in the frequency range above $5f_b$. A calculation of this effect shows it to be of the order of 5°. The effective load of the first two stages is about one-tenth of the load impedance of the last stage, so that the phase change introduced by the capacitance C_c will be negligible in these stages. From the above considerations it is seen that a phase margin of 15° would just meet the above requirements, and if this is doubled to 30° a reasonable margin of safety would be obtained.

(5) EXAMPLES OF FEEDBACK AMPLIFIERS

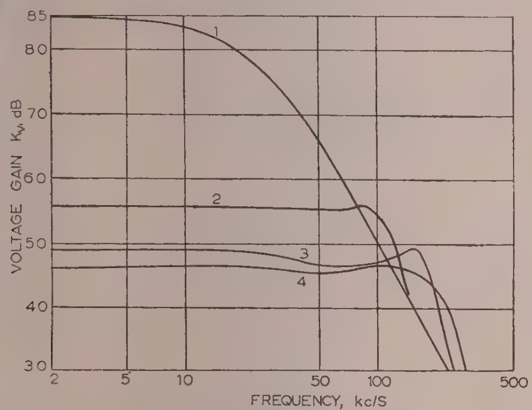
In Section 4 it was shown that the permissible limit of about 19 dB for purely resistive negative feedback could be increased considerably by the use of phase-compensating networks in the feedback circuit. In this Section, examples of feedback amplifiers are discussed in more detail.*

(5.1) Shunt Feedback Circuits

Three simple methods of producing phase compensation are described below. The three feedback networks employed are shown in Fig 6. The first, (*a*), gives the phase area plotted in

* The transistors employed in the feedback amplifiers were of *p–n–p* alloyed-junction type having $\alpha_0 \simeq 0{\cdot}95$ and $f\alpha \simeq 500$ kc/s.

7, curve (1), with $C = 1/2\pi R_2 f_b$ and the other elements of same values as in previous calculations. This curve is ed relative to the line of 180° of phase lag for the voltage or ent gain of the amplifier with no feedback. The point where e (1) crosses the curve A for the phase lag of the forward age-gain K_{v0} (or of current gain K_{i0}) is the point on the uency axis where the total phase angle, arg K_v, reaches 0°. At this point the current gain would be reduced by roximately 15 dB per stage, or a total of 45 dB, and the smission through the feedback circuit would be increased by roximately 15 dB. It is possible, therefore, to apply up to B of feedback with this circuit. There would, however, be margin of stability if the full 30 dB were applied. Curve (2) Fig. 8 shows the voltage-gain/frequency response obtained



ig. 8.—Voltage-gain/frequency response of experimental feedback amplifiers.

this type of circuit with $R_2 = 68$ kilohms and $C = 51\ \mu\mu$F, d curve (1) shows the response of the amplifier with no feed-k applied. The calculated Nyquist diagram for this circuit is wn in Fig. 9, curve (1). From this curve it can be seen that

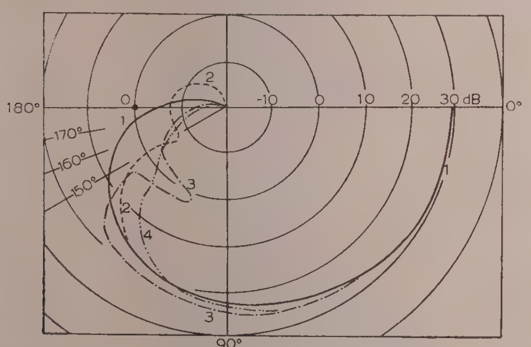


Fig. 9.—Nyquist diagrams for experimental feedback amplifiers. (Calculated complex loci of return ratio T.)

low-frequency value of $(-K_{v0}\beta)$ is approximately 29 dB. e margin of stability is very small.

f the amount of feedback needed is greater than 30 dB, or a sonable margin of stability is to be obtained, a more complex dback circuit must be used. Two suitable circuits are shown Figs. 6(*b*) and 6(*c*). Curve (2) in Fig. 7 shows the phase area ained with the circuit of Fig. 6(*b*) with $C = 1/6{\cdot}6\pi R_2 f_b$ and $= R_L/20\pi f_b$. Curve (2) crosses the original curve at $12 f_b$ on frequency axis, and at this point the total current gain is uced by more than 60 dB. The increase in the transmission through the feedback circuit is approximately 20 dB. The net effect on the loop gain will be a reduction of 40 dB at the frequency where the total angle has reached 180°. It is therefore possible to apply about 10 dB more feedback with this circuit than with that shown in Fig. 6(*a*).

A detailed analysis of the network shown in Fig. 6(*c*) is somewhat involved and will not be given here. Curve (3) of Fig. 7 shows the phase characteristic calculated for this type of circuit. The circuit would allow feedback of over 35 dB to be applied before the amplifier became unstable.

The circuits shown in Fig. 6(*b*) and 6(*c*) can be used, not to increase the amount of feedback beyond the original 29 dB obtained with the simple circuit, but to obtain an adequate margin of stability. Curve (2) of Fig. 9 shows the Nyquist diagram for the circuit illustrated in Fig. 6(*b*). This amplifier has the same low-frequency value of loop gain as the first example, but it is possible in this case to hold the phase angle of the loop gain to below 150° until the modulus has been reduced to $-7{\cdot}5$ dB. Curve (3) of Fig. 9 shows a typical Nyquist diagram for an amplifier using the feedback circuit shown in Fig. 6(*c*). This particular curve could probably have been chosen with a little more care, allowing the loop gain to approach closer to the 150° line at lower frequencies, and applying the rather excessive phase correction at a higher frequency. However, the amplifier has a phase margin of over 30° and a modulus margin of approximately 3 dB.

It was found in the calculations that using feedback of approximately 30 dB with the shunt-feedback amplifier would give considerable voltage-gain stability. To check this experimentally, a transistor of very low α (i.e. $\alpha_0 = 0{\cdot}89$) was interchanged with one of the transistors in a 3-stage amplifier that had good low-frequency α (i.e. $\alpha_0 = 0{\cdot}95$). The observed change of low-frequency gain for this interchange of transistors was approximately 9 dB when no feedback was applied, and less than $0{\cdot}5$ dB when negative feedback of 29 dB was applied. There was, however, considerable change in input impedance with this interchange of transistors. This change was found to be of the same order of magnitude as the change in voltage gain that was observed with no feedback.

(5.2) Series Feedback Circuits

Most of the previous discussion has related to amplifiers having shunt feedback. As pointed out previously, series feedback may be applied through a resistor in the common-emitter lead. If the series feedback is a pure resistance, there is the same limitation on the amount of feedback that can be applied to a 3-stage amplifier as for shunt feedback (i.e. approximately 19 dB). The phase-area theorem described in Section 4 can also be applied to this circuit. In the series feedback amplifier, the high- and low-frequency transmission is controlled by the series feedback element R_1 in parallel with the series combination of the generator impedance and the input impedance of the first stage. The difference in transmission at high and low frequencies can be made a maximum by arranging the series feedback element to be effectively open-circuited at higher frequencies. For the same amount of feedback at low frequencies there is more phase area available for phase compensation in the series than in the shunt feedback amplifier. However, although more phase area is available, the series network is a little more limited in its application, since it is necessarily a 2-terminal one.

The simplest method of obtaining a large part of the available phase compensation is to insert a small inductance in series with the feedback element R_1. Curve (3) in Fig. 8 shows the voltage-gain/frequency response obtained with an amplifier using this type of feedback. From the response curve without feedback

[curve (1)] it can be seen that feedback of approximately 38 dB was used in this circuit, and the amplifier remained stable.

Although considerable increase in feedback can be obtained with this simple circuit, the feedback must be greatly reduced to obtain a 30° phase margin. As pointed out previously, the forward gain without feedback acts very similarly to three simple minimum-phase networks, each having a modulus asymptote of 6 dB per octave, and the associated phase shift. The introduction of the inductance in series with R_1 cancels out the effect of one of the stages, as far as the loop gain is concerned, over most of the frequency interval of interest. The loop gain, therefore, has a modulus asymptote of 12 dB per octave, with the same phase angle as would be associated with a minimum-phase network of this slope. The modulus response of such a network is reduced by approximately 22 dB from its low-frequency value when the total phase angle is 150°. Thus the maximum amount of feedback that can be applied with this circuit, a phase margin of 30° being retained, is approximately 22 dB. The necessary modulus margin would have to be subtracted. This figure can be considerably increased by the use of a more complex feedback circuit. The problem is one of shaping the asymptote of the loop gain. It has been shown[3] that when a phase margin of 30° is required, the optimum slope is in the region of 10 dB per octave. The transmission through the feedback circuit must be made to increase by about 8 dB per octave to reduce the total slope of the loop-gain curve to 10 dB per octave. This can be achieved, with only slight modification of the simple circuit, if the transmission through the feedback circuit is allowed to decrease slightly at first and then increase more rapidly in the region where the 10 dB slope is needed. The modification of the circuit consists in connecting a condenser across the resistor R_1;

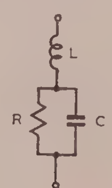


Fig. 10.—Series feedback network giving phase advance.

the feedback network is then as shown in Fig. 10. The inductance L and the condenser C have negligible effect in the useful frequency range of the amplifier. The Nyquist diagram for an amplifier using such a feedback circuit is shown in Fig. 9, curve (4). For this particular amplifier L and C are chosen such that $R_1C = L/R_1 = 1/2\pi f_0$ where $f_0 = 1\cdot67 f_b$. It can be seen from this curve that with 30 dB loop gain at low frequencies the amplifier would have a 30° phase margin and over 3 dB modulus margin.

With an interchange of transistors, the series-feedback amplifier gave the same degree of voltage-gain stability for the same amount of feedback as did the shunt-feedback amplifier. The series-feedback amplifier still has the disadvantage, however, that the input impedances are dependent on the current gain of each transistor. In specific applications, where the input and output impedances have strict limitations regarding return losses, fairly constant impedances could perhaps be obtained by including elements external to the transistor circuits, thus dictating the impedance level of the amplifier. The approximate formulae in Table 1 show that with compound feedback, where both series and shunt feedback are used on the same amplifier, the input and output impedances become to a large extent independent of the transistors.

(5.3) Compound Feedback

When both series and shunt feedback are applied to the same amplifier, the limitations on the total amount of feedback that can be applied without any phase compensation are the same as when either feedback is applied separately. Curve (4) in Fig. 8 shows the frequency response obtained for an amplifier using approximately equal amounts of shunt and series feedback. It can be seen by comparing this curve with curve (1) that the total amount of feedback in the mid-frequency range is approximately 36 dB.

It was found that with compound feedback amounting to 30 dB the change in gain by interchanging transistors was less than 0·5 dB, where the same interchange without feedback produced up to 9 dB change in voltage gain. Although no accurate measurements were made on the input and output impedances of the amplifier, rough measurements showed that the impedances stayed within 5% of their original value for any interchange of transistors of the same type, and remained constant with frequency up to approximately f_b.

The design of feedback amplifiers using both series and shunt feedback is basically the same as when both types of feedback are applied separately. In most cases of compound feedback the phase compensation for the series and shunt elements can be considered as a separate problem, and if they both satisfy the phase margin requirement the total loop gain will also satisfy it. It is then only necessary to make sure that the total loop gain satisfies the modulus margin. If the series and shunt loop gains are to be the same, this is equivalent to extending the modulus margins for the loop gains considered separately by 6 dB.

(5.4) Overall Power Gain

In the above examples only the overall voltage gain K_v of the amplifier has been referred to. From eqn. (5), however, the overall operating power gain is very simply related to K_g by the expression

$$G_0 = K_v^2 \frac{4R_g}{R_L}$$

The same values of R_g and R_L, namely 100 ohms and 5 kilohms, have been used in all examples, so that in decibels

$$(G_0)_{dB} = 20 \log_{10} K_v + 10 \log_{10} (400/5\,000)$$
$$= (K_v)_{dB} + 11\,\mathrm{dB}$$

(6) CONCLUSION

It has been demonstrated that useful negative feedback can be applied to a 3-stage common-emitter amplifier with a reasonable margin of dynamic stability, making its performance to a large degree independent of small changes in transistor parameters. Using series, shunt and compound feedback, and quite simple feedback networks, an overall operating power gain of about 33 dB with negative feedback of more than 30 dB over a frequency range of at least 20 kc/s has been achieved. Examples of amplifiers with such performance have been given, employing junction transistors with cut-off frequency f_α of approximately 500 kc/s. The overall gain characteristics obtained are substantially flat over a frequency range of about 100 kc/s, although the full negative feedback is of course applied over only a fraction of this range. From reports of the dependability of transistors when used in a circuit operated uninterruptedly, the gain stability achieved would seem adequate for many practical applications. It should be emphasized, however, that there are drawbacks, not discussed in the paper, with transistors at present available which limit their practical application. The most important of these are the limited frequency range (low f_α) and small power output

(7) ACKNOWLEDGMENTS

e authors wish to acknowledge the assistance of the Depart- of Scientific and Industrial Research, and of Marconi's less Telegraph Company, Ltd., for providing equipment for in transistor-circuit investigations. They are also very ful to Mullard, Ltd., for donating junction transistors of ype used in the circuits described in the paper at a time when ion transistors were otherwise unobtainable in this country.

(8) REFERENCES

(1) SHEA, R. F.: "Principles of Transistor Circuits" (Wiley, New York, 1953).
(2) THOMAS, D. E.: "Transistor Amplifier Cut-off Frequency," *Proceedings of the Institute of Radio Engineers*, 1952, **40**, p. 1481.
(3) BODE, H. W.: "Network Analysis and Feedback Amplifier Design" (Van Nostrand, 1945).

DISCUSSION ON

ECHNICAL ARRANGEMENTS FOR THE SOUND AND TELEVISION BROADCASTS OF THE CORONATION CEREMONIES ON 2ND JUNE, 1953"*

Before the WESTERN CENTRE *at* CARDIFF *8th February, the* SOUTHERN CENTRE *at* HOVE *3rd March, the* NORTH MIDLAND CENTRE *at* LEEDS *2nd November*, and the SOUTH MIDLAND CENTRE *at* BIRMINGHAM *6th December*, 1954.

Mr. F. J. Haines (*at Cardiff*): Were there pre-fade facilities on microphones at Westminster Abbey, or was all the fading formed "blind"?

Mr. J. Vaughan Harries (*at Cardiff*): I note that convertors re used in Paris and Breda to change the British 405-line ture to continental 625- and 819-line standards. Is this a manent installation, or was it installed only for the Corona- n? I assume that the system devised arose because no other thods of conversion are available. I should like the authors' ws on the picture depreciation brought about by this method.

Mr. E. K. Dalby (*at Hove*): I was interested in the reference the frequency response of the microphones used in the Abbey the means adopted for acoustic correction, but feel that, ticularly where surrounding objects affected the response, its racteristics could be attributed to the acoustics rather than he microphones. The information regarding the effectiveness he gauze screen in the lip microphone as an attenuator of the er frequencies provokes interest. Is it really an acoustic stance, discriminating and dissipating lower-frequency energy riction, or does it function as an acoustic impedance producing cellation of the lower frequencies by phase displacement? is well known, sound waves in the vicinity of a point source spherical with an inherent tendency to "bass lift." Could screen accelerate their conversion to planar waves?

he tapping of the amplifiers, through trap valves, at various ges and at suitable levels for particular requirements is of erest. What is a trap valve, and is it more generally known a cathode-follower? This device is ideal for such a purpose, ere a low-impedance feeder is to be connected to a high- edance system without adverse effects.

What type of device was used to amplify the centrimetre-wave quencies. Was the travelling-wave tube sufficiently developed he time? As a continuous-wave, virtually untuned, amplifier ould appear to be without rival, and must eventually supplant klystron in many applications.

Ve in the Eastbourne area have been neglected by the B.B.C., ough the phenomenal and consistent television signal strength Coronation Day exceeded all expectations. Could this be to an increase in effective radiated power?

Since the announcement that reliance was being placed upon aerial gain for increased e.r.p. from the new Crystal Palace transmitter, could not the B.B.C. change the existing Alexandra Palace aerial, which has a gain of 2 and so an e.r.p. of 34 kW, for a Crystal Palace type of aerial, with a gain of 8 and so increase the e.r.p. to 136 kW? The resultant reduction of the vertical angle of the 360° fan beam would also reduce the troublesome reflections from aircraft, etc. I consider that, compared with the benefits conferred upon all Alexandra Palace fringe viewers, the cost of this change would be negligible; it should prove cheaper, more equitable and effective than the present policy of installing booster stations in selected areas, which improve conditions for a favoured minority only.

Dr. G. N. Patchett (*at Leeds*): It seems surprising that it was possible to locate four television cameras in Westminster Abbey so that they were invisible and yet able to give such excellent pictures of the Ceremony. It is also interesting to note that the large amount of extra lighting required by the film cameras, particularly for colour, is much greater than that required for television.

Were any special precautions taken against vibration and wind when the long-focus lens, giving only a 9 ft coverage at 300 ft, was used?

Are the derivative equalizers adjusted on a test card or on an actual scene; and are they now used as a normal procedure?

The use of a picture-synchronizing signal from a master oscillator was ingenious: what is the maximum angle before frame slip occurs?

Mr. L. L. Tolley (*at Birmingham*): Thorough arrangements for intercommunication were obviously necessary for such a large-scale project, but I presume that the producer was adequately safeguarded from telephone interruptions. The paper mentions an arrangement by which the producer can take a cue line and a contribution line and thereby get two-way conversation with his commentators. I imagine that this was under his control and that if the commentators wanted to speak to the producer they would use the telephone circuit which would be answered by an operator, to save the producer from undue interruption?

It was an excellent idea to put the record on the long-distance lines. There are probably hundreds of circuits passing through

PROCTER, W. S., PULLING, M. J. L., and WILLIAMS, F.: Paper No. 1593 R, ember, 1953 (see **101**, Part I, p. 57).

each repeater station on the route, and it is very easy for the operator to misread his records, to miscount or to do something which brings him to an important circuit in the belief that it is a spare not in use. He goes on the lines to confirm that there is nothing there, and if there is silence it confirms his belief.

Fig. 9 shows the arrangement of the circuits from the cameras to Broadcasting House, and one notes that appreciable parts of the video links were routed over telephone circuits, as is not uncommon. Telephone cable is not designed for television frequencies, but the video amplifiers inserted in the links overcame the problem from the attenuation aspect; however, did any troubles occur owing to noise or reflections? Fig. 9 shows ten video repeaters as main repeaters as well as three spares, but the paper says that seven were in simultaneous use. Is there a contradiction here or have I misinterpreted the word "simultaneously"?

To synchronize the cameras the line frequency of 10 125 c/s was transmitted over the telephone cable. This is beyond the range of frequencies normally used for telephone speech, and I wonder whether there was any trouble due to crosstalk.

I note that the microphone leads in the Abbey were very long. I imagine that it was impossible to carry out the first stage of amplification near the microphones; was there any trouble from induction in the long leads? They were, of course, some distance from the high-power lighting circuits.

Mr. F. G. Furniss (*at Birmingham*): I was interested in the large number of microphones installed in the Abbey for which various compensations had to be undertaken. Were the compensations carried out individually, and was the effect of the frequency distortion corrected by aural test rather than by measurement? Very many authorities had to use the Abbey prior to the Coronation, and all required a lot of time. A great deal of time must also have been spent in organization, testing and improving the installation. How was all that co-ordinated, and who undertook it?

Dr. J. H. Walker (*at Birmingham*): The authors comment on the high cost of hiring Post Office lines. It would be interesting to know how the general question of cost is dealt with, in particular that of the services provided for the Continent and North America; were they supplied on a commercial basis or was there a reciprocal agreement concerning future programmes relayed from the Continent?

Mr. W. T. Gemmell (*at Birmingham*): If the authors had to undertake this task again would they make any changes? The general impression in retrospect is that it could not have been better, but the authors have been careful to say that they were merely applying known techniques on a very much larger scale than before, and that success was a matter of organization, collaboration and co-operation between all parties.

Mr. G. O. Romans also contributed to the discussion at Cardiff and **Dr. K. R. Sturley** to the discussion at Birmingham.

Messrs. W. S. Procter, M. J. L. Pulling and **F. Williams** (*in reply*): In reply to Mr. Haines, there were no pre-fade facilities on the microphones in Westminster Abbey. The mixing was carried out by an expert engineer who had, in fact, performed the same task at the previous Coronation broadcast. He had, of course, had the opportunity of taking part in all the rehearsals and had familiarized himself with every detail of the Ceremony.

In reply to Mr. Vaughan Harries, the standards convertors installed at Paris and Breda were, at the time of the broadcast, temporary installations. It is quite correct that the method of standards conversion described in the paper was the only one available at the time—indeed, this is still the case. Some degradation of picture quality is inevitable, but this is less serious when converting to a lower-definition picture than when converting in the opposite direction.

Replying to Mr. Dalby, the gauze screens in the lip microphone function as acoustic resistances. They are placed close to the surface of the ribbon and impose a load which at frequencies below 1 kc/s exceeds the existing mechanical impedance of the latter. Under these conditions the ribbon, being resistance controlled, generates a voltage which is proportional to frequency.

The function of a trap valve is simply to enable a number of circuits to be fed independently from a single source in such a way that, if a fault should develop in any one of the circuits being fed, there is no effect in the feeds to the others. Trap valves are not used primarily as impedance-changing devices.

Both klystrons and travelling-wave tubes were in use in centimetric-wave equipments of different types employed on this broadcast.

The power radiated by B.B.C. television transmitters on Coronation Day was in no way abnormal. Many viewers reported apparently improved signal strength, but this is thought to have been due in the majority of cases to the fact that a great deal of normal work was at a standstill and car traffic at a low level. These factors contributed to a generally very low level of electrical interference and, consequently, to an improvement in the average signal/noise ratio.

In reply to Dr. Patchett, considerable care has to be taken when using long-focal-length lenses to ensure that cameras are firmly mounted, and that their supporting tripods and any scaffolding on which they may be mounted are as rigid as possible.

Derivative equalizers are adjusted on a test card and are now used as part of the normal camera equipment.

We think that Dr. Patchett may have misunderstood the description of the arrangement for synchronizing a number of independent camera sources by means of a master oscillator. When this arrangement is used the individual pulse generators are kept absolutely in synchronism by means of the signal from the master oscillator.

Replying to Mr. Tolley, commentators were definitely discouraged from calling the producer during transmission. If a commentator needed to do so, he would use the telephone with which he was provided. A hand-ringer on this telephone would cause an indicator to drop on the telephone switchboard next to the producer's position. The telephone would be answered by an engineer, who would ask the producer to take the call at an appropriate moment. In the opposite direction, the producer would call the commentator by means of his microphone and talk-back circuit, the commentator hearing him through the headphones which he would be wearing.

No trouble was experienced due to noise or reflections on telephone circuits used to carry video signals; nor was there any trouble due to crosstalk on the telephone circuits carrying the 10 125 c/s synchronizing tone.

Mr. Tolley is quite correct in assuming that, in the Abbey, there were no microphone amplifiers near the microphones themselves. By taking the normal precautions of individually screening each microphone lead from the microphone to the control room, and by careful earthing of the screens, induction troubles in these long circuits were avoided.

In reply to Mr. Furniss, compensations of the microphone circuits in the Abbey were effected individually and purely by the aural assessment of expert engineers.

In reply to Dr. Walker, the circuits provided by the General Post Office for sending programmes to the Continent and elsewhere were supplied on a perfectly normal commercial basis.

In reply to Mr. Gemmell, we think that if a similar job had to be done again, very little significant change would be made, except possibly on the television side where, if we had more camera equipment available, we would hope to be able to improve the "coverage," especially on the processional route.

6.677.71

Paper No. 1930 R
Jan. 1956

INVESTIGATION OF SLOT RADIATORS IN RECTANGULAR METAL PLATES

By D. G. FROOD, B.A., M.A., and J. R. WAIT, M.Sc., Ph.D.

(*The paper was first received* 16*th May, and in revised form* 2*nd August*, 1955.)

SUMMARY

e radiation from slots cut in conducting surfaces of limited extent cussed. Equatorial plane patterns of an axial half-wave slot in a ngular metal plate are measured in the X band. The experimental s compare favourably with the calculated patterns on the assump- that the plate can be represented by a thin elliptic cylinder or n of infinite length. It is observed that, if the length of the plate ıal to or greater than its width, the pattern is within a few per cent e corresponding theoretical pattern for a plate of infinite length. e admittance of the slot in the plate was also measured and com- d with the computed conductance. The agreement is seen to be good.

LIST OF PRINCIPAL SYMBOLS

$\phi, z)$ = Radial, azimuthal and axial co-ordinates of a cylindrical co-ordinate system.
V = Voltage across the centre of the slot.
η_0 = Intrinsic impedance of free space (=120 ohms).
$k = 2\pi/\lambda$.
r = Distance from observer to centre of slot.
r_1 = Distance from observer to lower end of slot.
r_2 = Distance from observer to upper end of slot.
ω = Angular frequency.
θ = Polar angle = arc tan z/ρ.
$S(\theta)$ = H-plane pattern of slot on infinite sheet.
Y = Self-admittance at centre of slot.
Z = Self-impedance of slot ($=1/Y$).
$\Phi, z)$ = Cylindrical co-ordinates centred at one edge of sheet for principal E-plane.
$\Phi', z)$ = Cylindrical co-ordinates centred at second edge of sheet for principal E-plane.
$F(s)$ = Fresnel-type integral with upper limit s defined in eqn. (6).
S_1, S_2 = Limits for the integral $F(s)$.
H_0 = Axial magnetic field of incident plane wave.
K = Factor of proportionality depending on the geometry of the slot.
v = Voltage induced at the centre of the slot by the incident wave.
$g(kd)$ = Shunt conductance of the slot in the waveguide divided by the characteristic admittance of the waveguide (= function of kd).
a, b = Inner dimensions of the broad and narrow faces of the waveguide.
λ_g = Effective wavelength in guide.

(1) INTRODUCTION

lot antennae are becoming very extensively employed in rowave radiating systems. Their history of development has n rapid. It is difficult to say exactly when they were invented, certainly the contributions from Watson[1] and Booker[2] and r collaborators were of major importance.

It is surprising that little attention has been paid by previous workers to the effect of cutting the slot in metal surfaces of finite extent. While the developed design procedures assumed that the exterior surface was infinite in extent, the need for a practical antenna system limits the physical size of the metal surface on which the slots are cut. Stevenson,[3] who developed an elegant theory for the radiation of resonant slots in a rectangular waveguide, assumed for convenience that the exterior region was equivalent to a half space. He admits that the assumed infinite dimension of the face of the waveguide for the exterior problem is a severe limitation to the validity of the expressions for the conductance of the slot.

It is the purpose of the paper to investigate in some detail the significance of the finite extent of the metal surface or sheet on which the slot is cut.

(2) THEORETICAL DISCUSSION

The radiation pattern of a thin slot cut in an infinite plane sheet of perfect conductivity and vanishing thickness can be obtained by an application of an electromagnetic Babinet's principle.[2] By using this technique, the well-known results[4] for the thin-wire antenna can be transformed immediately to the complementary problem. For a half-wave slot oriented in the z-direction of a (ρ, ϕ, z) co-ordinate system and centre-fed by a voltage V, the fields are given by

$$\left.\begin{aligned} H_z &= \frac{-jV}{2\pi\eta_0}\left(\frac{\varepsilon^{-jkr_1}}{r_1} + \frac{\varepsilon^{-jkr_2}}{r_2}\right)\varepsilon^{j\omega t} \\ H_\rho &= \frac{jV}{2\pi\eta_0\rho}\left[(z+\lambda/4)\frac{\varepsilon^{-jkr_1}}{r_1} + (z-\lambda/4)\frac{\varepsilon^{-jkr_2}}{r_2}\right]\varepsilon^{j\omega t} \\ E_\phi &= \frac{-jV}{2\pi\rho}(\varepsilon^{-jkr_1} + \varepsilon^{-jkr_2})\varepsilon^{j\omega t} \end{aligned}\right\} \quad . \quad (1)$$

where r_2, r_1 and r are the distances of any point P to the upper, centre and lower end, respectively, of the slot and where $\eta_0 = 120\pi$ ohms and $k = 2\pi/$(free-space wavelength). Eqns. (1) refer to the slot cut in an infinite sheet. When the point P is sufficiently far away from the slot the equations simplify somewhat to

$$\left.\begin{aligned} E_\phi &\simeq j(2\pi r)^{-1}S(\theta)\exp(-jkr + j\omega t) \\ H_z &\simeq \left(\frac{1}{\eta_0}\right)E_\phi \sin\theta \\ H_\rho &\simeq -\left(\frac{1}{\eta_0}\right)E_\phi \cos\theta \end{aligned}\right\} \quad . \quad . \quad (2)$$

where $\theta = \text{arc tan}\,(z/\rho)$ and where terms containing r^{-2}, r^{-3}, etc., have been neglected. The factor $S(\theta)$ can be defined as the radiation pattern of the slot and is given by

$$S(\theta) = V\frac{\cos\left(\frac{\pi}{2}\cos\theta\right)}{\sin\theta} \quad . \quad . \quad . \quad . \quad . \quad (3)$$

Following the method of Carter[5] for the thin-wire antenna,

ritten contributions on papers published without being read at meetings are ed for consideration with a view to publication.
r. Frood and Dr. Wait were formerly at the Defence Research Telecommunications blishment, Ontario, Canada.
r. Frood is in the Department of Theoretical Physics, University of Liverpool.
r. Wait is at the Central Propagation Laboratory, National Bureau of Standards, rado, U.S.A.

the self-admittance Y at the centre of the slot can be obtained from

$$Y = \frac{-2}{V}\int_{-\lambda/4}^{\lambda/4} H_z(0, 0, z) \sin kz dz \quad . \quad . \quad . \quad (4)$$

The integrations can be carried out to yield the result

$$Y = 2{\cdot}06 + j0{\cdot}97 \text{ millimhos}$$

which corresponds to the case when the slot is allowed to radiate on both sides of the sheet. The centre impedance of the slot is then

$$Z = 1/Y = 365 - j212 \text{ ohms}$$

It should be noted that Begobich[6] recently gave the value $362{\cdot}5 + j210{\cdot}5$ ohms, which indicates reasonable agreement with the above, except for the sign of the reactance. Apparently he made an error in his derivation. The result derived by the authors can be checked by comparing it with Booker's statement of Babinet's principle, and connecting the impedance, Z, of the slot with the impedance, Z^w, of the complementary wire antenna, such that

$$ZZ^w = \eta_0^2/4 = 3\,600\pi^2$$

Using Carter's[5] value $73{\cdot}2 + j42{\cdot}5$ for Z^w leads back to the authors' expression for Z. Begovich erroneously employed the complex conjugate of Z^w.

When the slot radiates only on one side of the sheet the conductance G would be one-half the real part of Y given above, and hence $G = 1{\cdot}03$ millimhos. The corresponding susceptance is one-half the imaginary part of Y plus the susceptance of the feed system.

It should be noted that the pattern of the slot on the infinite sheet is essentially omni-directional in the azimuthal plane. It is of interest to consider the effect of truncating the sheet. For example, the slot is cut in the centre of a rectangular metal plate of width $2d$, and oriented parallel to the other sides of length L,

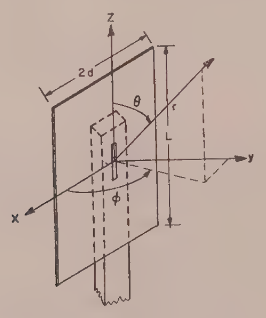


Fig. 1.—Thin half-wave slot cut in the centre of a rectangular metal sheet and co-ordinate system.

as illustrated in Fig. 1. The slot is fed by a waveguide such that radiation takes place only on one side of the plate.

The ideal E-plane pattern in the equatorial plane, corresponding to a value of d very much greater than the wavelength, would be as shown in Fig. 2, where the plane of the sheet corresponds to $\Phi = 0°$. In other words, the sheet is assumed to be sufficiently wide that diffraction around and by the edges is negligible. Of course, if the slot were allowed to radiate on both sides of the sheet the pattern would be a complete circle. The corresponding H-plane pattern for the half-wave slot is simply the function $S(\theta)$ shown in Fig. 3, where $\theta = 0°$ is in the plane of the sheet.

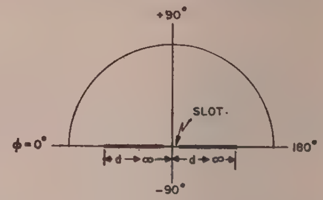


Fig. 2.—E-plane pattern of the slot for an infinitely large sheet.

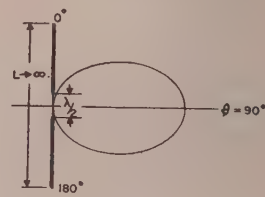


Fig. 3.—H-plane pattern of the slot for an infinitely large sheet.

It would now be expected, in view of the shape of the H-plane pattern shown in Fig. 3, that the effect of the finite value of L would not be pronounced. This is later borne out by experiment. However, the finite value of d can be expected to lead to a considerable impairment in the pattern, since the field in the broadside direction of the slot is very significant. It is possible to treat the problem of the slot in the finite plate by an approximate theoretical procedure, if the edges separated by the distance $2d$ are considered to diffract the primary field of the slot as if they were semi-infinite half-planes. The interaction between the edges is neglected, and on physical grounds this would seem justified if $2d$ is somewhat greater than the free-space wavelength. Again, this supposition is borne out by experiment and comparison with a more rigorous treatment, such as that which represents the plate by a thin elliptic cylinder.[7,8]

To facilitate the discussion it is desirable to consider the slot as a receiving element. The voltage induced in the thin slot from the wave incident on the plate is then proportional to the tangential magnetic field along the slot. This statement follows from Schelkunoff's equivalence principle.[4] From Fig. 4 a plane

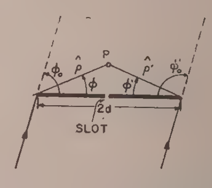


Fig. 4.—End view of the sheet showing a plane wave incident from below.

wave is shown to be incident from below the thin conducting plate of width $2d$. The incident wave is polarized such that the magnetic vector H_0 is parallel to both the slot and the edges. Choosing a polar co-ordinate system centred at the left-hand edge of Fig. 4, the field H_z at P of co-ordinates (ρ, Φ, z) due to diffraction around this edge is given by the classical Sommerfeld formula[9]

$$H_z = H_0[\varepsilon^{-jk\rho\cos(\Phi-\Phi_0)}F(s_1) + \varepsilon^{-jk\rho\cos(\Phi-\Phi_0)}F(s_2)] \quad . \quad (5)$$

where Φ_0 is the angle the incident waves makes with the plane of the sheet, where

$$\left.\begin{aligned} F(s) &= \varepsilon^{j\pi/4}\pi^{-1/2}\int_{-\infty}^{s}\varepsilon^{-jx^2}dx \\ \text{and} \qquad s_1 &= (2k\hat{\rho})^{1/2}\sin\left(\frac{\Phi-\Phi_0}{2}\right) \\ s_2 &= -(2k\hat{\rho})^{1/2}\sin\left(\frac{\Phi+\Phi_0}{2}\right) \end{aligned}\right\} \quad . \quad . \quad . \quad (6)$$

Similarly the field H_z' at P due to diffraction around the other or right-hand edge is identical in form to eqn. (5) if ρ, Φ and Φ_0 are replaced by $\hat{\rho}'$, Φ' and Φ_0, and H_0 is replaced by $H_0\varepsilon^{-j2kd\cos\Phi_0}$. The voltage v induced in the centrally located slot on the sheet is then given by

$$v = K(H_z + H_z') \text{ for } r' = r = d \text{ and } \Phi = \Phi' = 0$$

where K is a constant which depends on the dimensions of the slot. It then follows that

$$|v| = K\left|F[-(2kd)^{1/2}\sin(\Phi_0/2)] + F[-(2kd)^{1/2}\cos(\Phi_0/2)]\right| \quad . \quad . \quad . \quad . \quad (7)$$

which is applicable in the range $\Phi_0 = 0°–180°$.

When the incident wave is incident on the upper side of the sheet, steps must be taken to combine the fields, as determined by the two knife-edge problems, in the proper manner. It is convenient to choose a slightly different co-ordinate system as shown in Fig. 5, where the angular co-ordinates Φ and Φ_0 are now measured from the bottom of the sheet.

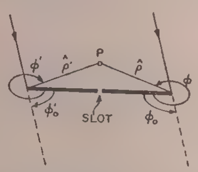


Fig. 5.—End view of the sheet showing a plane wave incident from above.

The magnetic field H_z at P, regarding the sheet as a semi-infinite half-plane with the edge at (0, 0, z), is given by eqn. (5), and a similar expression is obtained for H_z'. The voltage v induced in the slot is then given approximately by

$$v = K[H_z + H_z' - H_0\varepsilon^{jkd\cos\Phi_0}] \quad . \quad . \quad . \quad (8)$$

The term $H_0\varepsilon^{jkd\cos\Phi_0}$ is the primary field of the incident wave at the slot and it must be subtracted from $H_z + H_z'$, since the primary field is included both in H_z and H_z'.

The slot voltage for the wave incident on the upper side of the sheet is then given by

$$|v| = K\left|F[(2kd)^{1/2}\sin(\Phi_0/2)] + F[(2kd)^{1/2}\cos(\Phi_0/2)] - 1\right| \quad . \quad . \quad . \quad . \quad (9)$$

for Φ_0 in the range 0°–180°.

As a numerical example, the E-plane pattern for a thin slot cut in the axial direction on a metal sheet of width $2d$ is computed using eqns. (7) and (9) for $kd = 141$. The pattern plotted in Fig. 6 is normalized in the broadside direction to 0 dB. It is interesting to note that the field in the direction tangential to the sheet is 6 dB below the maximum value. This is a characteristic which holds consistently for all sheet widths greater than a few wavelengths. It is worth mentioning that Booker has given a value of 3 dB rather than 6 dB for the reduction of the field along the sheet. His argument is based on physical grounds, which apparently does not account properly for the energy diffracted around the edges. The calculations were also carried out for $kd = 94$ and 234, and these are shown plotted in Fig. 7 in the interesting transition region for angles within 15° above and below the plane of the sheet. It is noted that the field at the

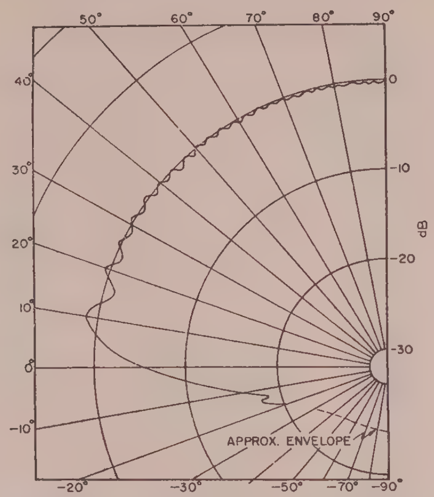


Fig. 6.—Principal E-plane pattern of a slot in a sheet of width $2d$ symmetrical about 90°. The plane of the sheet corresponds to 0°.

$kd = 141$
$\theta = 90°$
$V_0 = 90°$
N.B.—Curve is calculated.

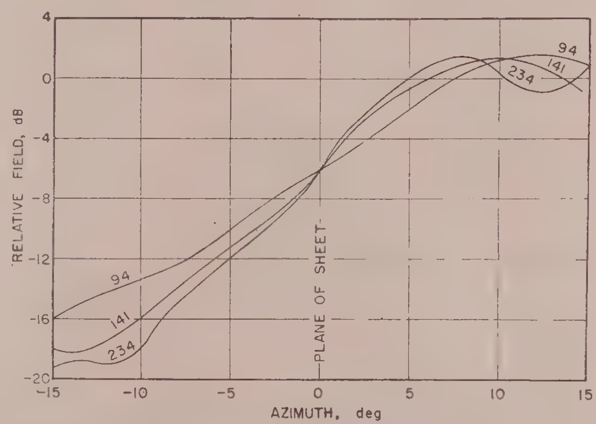


Fig. 7.—Behaviour of the E-plane pattern for directions 15° above and below the plane of the sheet.

Value of kd is indicated on the curves.
N.B.—Curves are calculated.

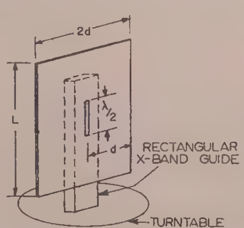
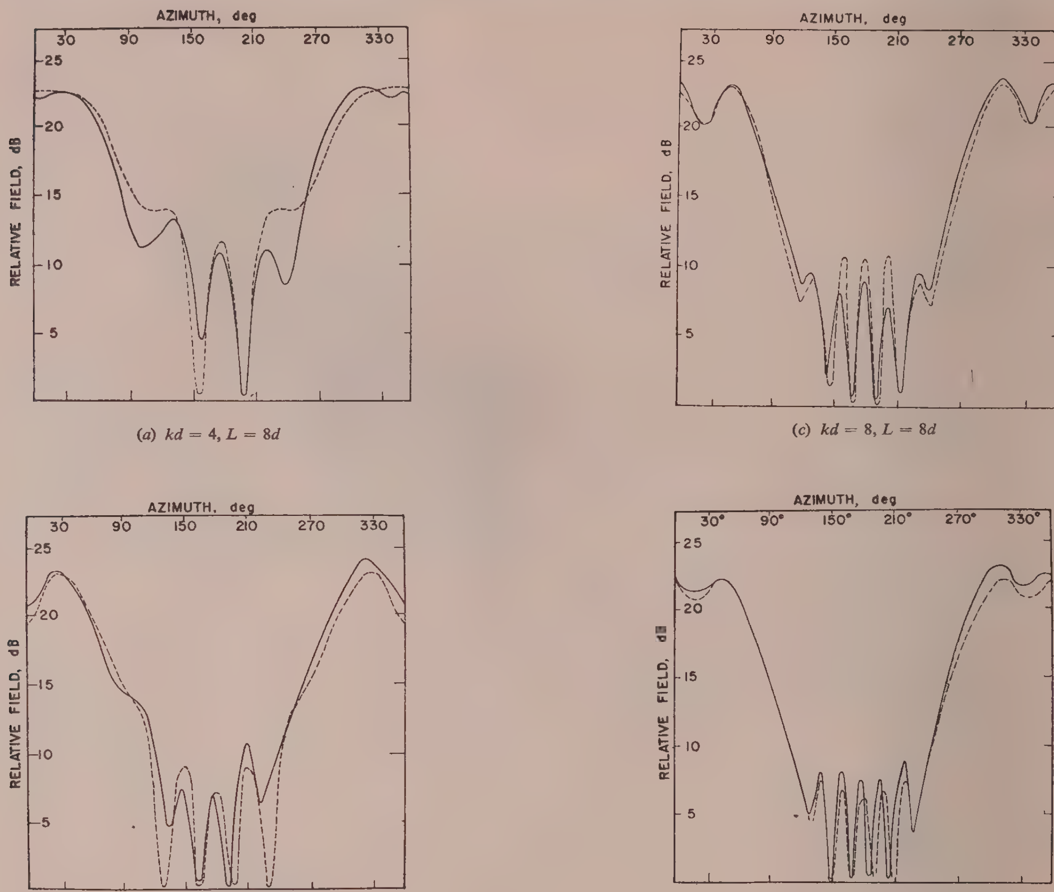


Fig. 8.—Schematic of the waveguide-fed slot mounted on the turntable for the pattern measurements.



(a) $kd = 4, L = 8d$ (c) $kd = 8, L = 8d$

(b) $kd = 6, L = 8d$ (d) $kd = 10, L = 8d$

Fig. 9.—Principal E-plane patterns of a thin half-wave slot in the centre of a rectangular sheet of length L and width $2d$.

——— Experimental.
---- Theoretical.
$\theta = 90°$
$V_0 = 90°$
Plate 4·7 ft wide at X band.

rear of the sheet diminishes very slowly with increasing values of $2d$. In fact, it is easy to show from the asymptotic behaviour of the integral F(s), that the amplitudes of the envelopes of the back lobes are approximately inversely proportional to $\sqrt{2d}$.

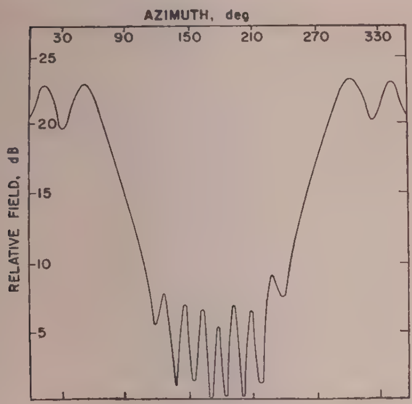
The above approximate method for calculating patterns of slots in metal sheets of finite width cannot be expected to yield reliable results when $2d$ is of the order of a wavelength or less. A more rigorous approach is to represent the sheet by a thin elliptic cylinder of vanishing minor axis. Computations based on this model have been carried out previously[8] for values of kd ranging from 2 to 8. It was shown that the double knife-edge approximation is accurate to within a few decibels if kd is greater than about 6.

(3) EXPERIMENTAL PATTERNS

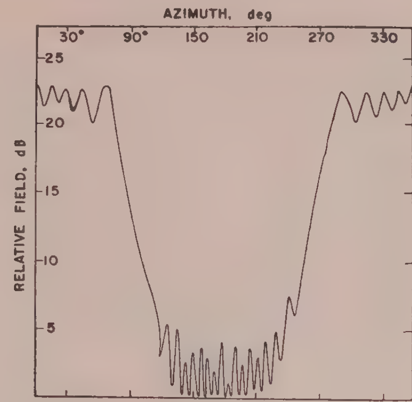
In this Section some experimental patterns will be compared with the patterns computed both from the double knife-edge technique and from the elliptic-cylinder method.

The experimental work was carried out on an antenna range at a wavelength of 3·2 cm. A narrow half-wave slot was cut in the broad face of an X-band waveguide illustrated in Fig. 8, with the centre of the slot approximately three-quarters of a guide wavelength from the short-circuited end. The slot was parallel to the axis of the guide and was offset approximately 0·1 in from the centre of the broad face of the guide. A number of thin rectangular aluminium plates were prepared with kd varying from 4 to 141 with length/width ratios from $\frac{1}{2}$ to 20. These plates were designed to be easily mounted on the broad face of the guide. The waveguide assembly was then mounted on a suitable turntable and illuminated by a transmitting dish antenna which was located at a distance of 100 ft. The radiation was incident normally to the axis of the guide and was horizontally polarized. The output from the waveguide was detected and amplified and its varying values were plotted on an ink recorder which had a logarithmic scale.

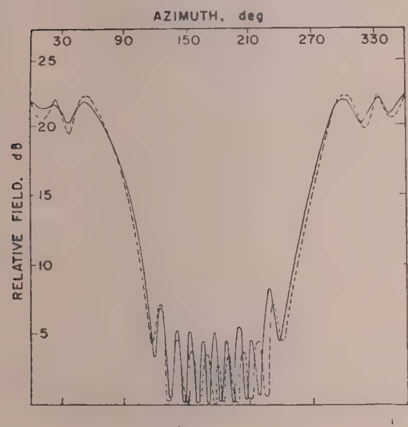
Measured patterns are shown in Figs. 9(a)–9(h) for $kd = 4$, 6, 8, 10, 12, 16, 28 and 141, for the case where the slot is centrally located in the plate. Theoretical data are also shown for comparison in some of the curves. The theoretical curve for the case $kd = 141$ is shown in Fig. 6. The vertical scale is arbitrary, and for the sake of convenience the experimental and calculated curves are matched in the direction of the maximum field. The elliptic-cylinder method of calculation was employed for kd values of up to 10 using eqn. (8) or Reference 8 with $\theta = 90°$,



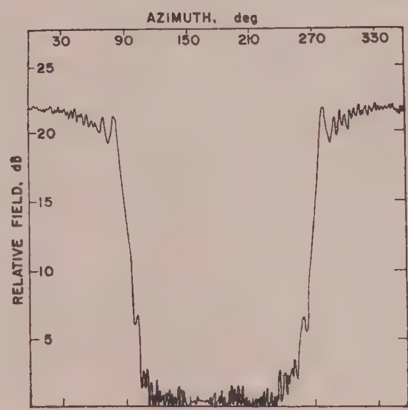
(e) $kd = 12, L = 8d$



(g) $kd = 28, L = d$



(f) $kd = 16, L = 8d$



(h) $kd = 141, L = 3d/2$

Fig. 9—*continued*

ıd v_0, the angular co-ordinate of the slot, equal to 90°. For rger values of kd the double knife-edge technique was used to otain the theoretical curves. It should be noted that the imuth scale of these figures are chosen so that the broadside rection from the slot corresponds to 0°.

The agreement between experiment and theory improves for e larger sheets. The probable reason for the discrepancy for nall sheets is that the diffraction by the waveguide behind the eet is becoming significant. For $kd = 4$ the plate is only oout 50% wider than the broad face of the guide, and it is erefore not surprising that the pattern differs from that calılated on the basis of a thin elliptic cylinder. It is also interestg to observe that the experimental pattern for $kd = 4$ is mewhat asymmetrical; this is due, no doubt, to the fact that hile the slot is centrally located in the plate, the guide is disaced slightly. The asymmetry is also seen to occur for $kd = 6$, ıt to a lesser extent.

To illustrate the effect of the length of the plate, patterns were corded for $L = d$, $4d$ and $20d$, keeping kd constant at 6. The milarity between these curves, shown in Fig. 10, is striking, ıd substantiates the earlier supposition that the azimuthal itterns are determined mainly by the lateral dimension (i.e. $2d$) the plate for an axial slot. In fact, it can be seen from the ırves in Fig. 10 that the plate can be regarded as infinite in ngth so long as L is greater than about $4d$. The main effect of finite length seems to be an increase in the level of the back lobes.

In all the above-mentioned experiments the slot was situated in the centre of the plate. If the slot was displaced toward one edge by an amount $d/2$, the angular elliptic co-ordinate

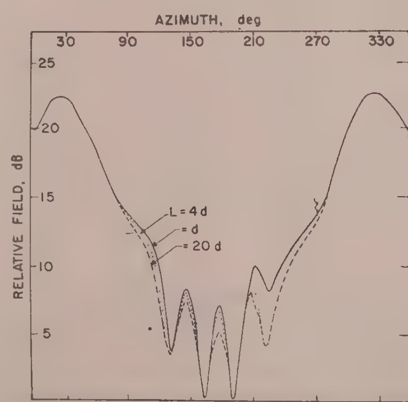


Fig. 10.—Principal E-plane pattern for the sheet of various lengths with constant width.

Effect of length $kd = 6$, $\theta = 90°$, $V_0 = 90°$, $L = d, 4d, 20d$.

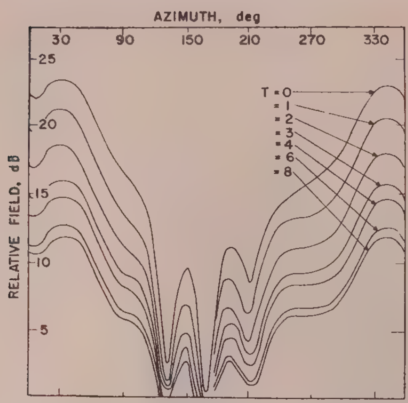


Fig. 11.—Principal *E*-plane patterns for the slot displaced toward the right-hand edge by an amount $d/2$. (The values of T correspond to the number of layers of tape covering the slot.)
$kd = 4$, $\theta = 90°$, $V_0 = 60°$.

becomes $v_0 = 60°$. The pattern then became quite asymmetrical as shown in Fig. 11. This type of asymmetry was also present in the theoretical computations of the offset slot or the thin elliptic cylinder.[8] The other curves in Fig. 11 correspond to the patterns for the slot covered with layers of plastic electrical tape of thickness 7·5 mils. The number of layers is indicated by the value T. It is interesting that the azimuthal patterns are essentially unchanged in shape, which is in accord with theory for a thin slot. It is not possible to draw any further conclusions from this set of curves, since it was not feasible at the time to measure the change of the voltage standing-wave ratio in the guide for the different thicknesses of the dielectric covering. Furthermore, the dielectric properties of the plastic tape are not known.

(4) MEASUREMENT OF SLOT ADMITTANCE

Another important characteristic of a slot cut in a metal surface is its admittance. In the earlier part of the paper it was shown that the admittance at the centre of a thin half-wave slot, cut in an infinitely thin conducting sheet of infinite extent, was $2·06 + j0·97$ millimhos. If the slot is fed by a waveguide, which is located on one side of the sheet so that it radiates only into one of the half-spaces, the conductance at the centre of the slot is 1·03 millimhos and the susceptance, as mentioned earlier, is dependent on the nature of the evanescent structure of the field within the guide.

When the sheet is of finite size the conductance is no longer 1·03 millimhos. The variation of G with the width of the sheet was investigated theoretically by using a model of a thin axial half-wave slot at the centre of the broad face of a thin elliptic cylinder.[8] G was obtained explicitly by computing the power radiated from the slot for a specified voltage at the centre of the slot. It was shown that G was an oscillating function of the width $2d$, and it approached 1·03 millimhos as $2d$ approached infinity.

It is now worth while to examine the admittance using an experimental procedure. The slot of width 1/16 in and length 5/8 in was cut parallel to the narrow face of the X-band waveguide. Means were then taken to mount a series of plates of various widths flush with the narrow face of the waveguide in a similar manner to that employed for the pattern measurements of the slots. The slot was cut in the narrow face rather than the broad face, so that the effect of plates of small width could be examined also.

As is customary in waveguide measurements, the longitudinal slot is represented by an equivalent shunt admittance across the equivalent transmission line of the waveguide. Employing a slotted line in conjunction with an adjustable short-circuit termination in the guide beyond the slot, the conductance and susceptance are determined using a standard technique.[10] These values are normalized by dividing by the characteristic admittance of the guide, which is real. The normalized conductance denoted by g, and the normalized susceptance denoted by b_n, are plotted in Fig. 12 as a function of kd.

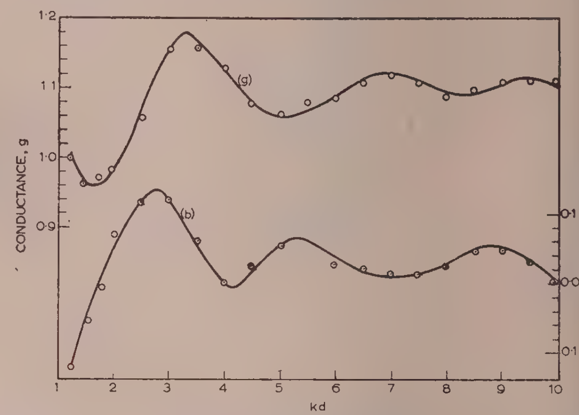


Fig. 12.—Measured normalized conductance and susceptance of the slot as a function of the width of the sheet.
Slot width = $\frac{1}{16}$ in
Slot length = $\frac{5}{8}$ in
$\lambda = 3·20$ cm, $\lambda_g = 4·48$ cm

By adapting Stevenson's theory[3] for longitudinal slots near resonance, cut in the narrow face of the waveguide, it is easy to show that the normalized conductance $g(kd)$, as a function of kd, in terms of the actual conductance G at the centre of the slot, is given by

$$g(kd) = \frac{480}{73\pi}\frac{a}{b}\frac{\lambda g}{\lambda}\cos^2\frac{\pi\lambda}{2\lambda g}\frac{1\cdot 03}{G} \quad . \quad . \quad . \quad (10)$$

Stevenson's formula corresponds to eqn. (10) when G is replaced by 1·03; this corresponds to the case of an infinite plate (i.e. $kd \to \infty$). Employing the theoretical results for G, mentioned above, the value of $g(kd) - g(\infty)$ given by eqn. (10) is plotted in Fig. 13 using $a = 0·90$ in, $b = 0·40$ in, $\lambda = 3·2$ cm, $\lambda_g = 4·48$ cm, along with the corresponding experimental curve.

There is reasonable agreement between the experimental and the

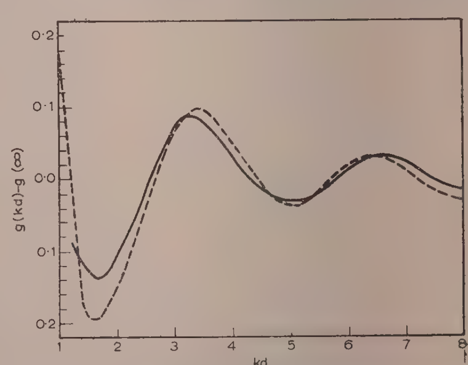


Fig. 13.—Comparison between the experimental and theoretical normalized conductance.
[N.B.—The theoretical value of $g(\infty)$ is 1·20.]
——— Experimental.
– – – – Theoretical.

mputed curves. The disagreement for the smaller values of kd n probably be accounted for by the fact that the theory does t account for the diffraction by the waveguide behind the plate. should also be mentioned that the theoretical value of g, hich is found using Stevenson's procedure, assumes that the sceptance is much smaller than the conductance (i.e. the slot near resonance). It can be seen from the experimental results at, although b_n is small, it is not negligible and would, no ubt, also be a source of discrepancy.

(5) CONCLUSIONS

It has been demonstrated that the radiation characteristics of axial half-wave slot in a rectangular metal plate are mainly a nction of the width, rather than the length, of the plate. The easured pattern and radiation conductance of the slot agreed ite closely with theory over a wide range of plate widths. ne experimental phase of this project is continuing, and parcular attention will be paid to the effect of the finite width d depth of the slot. It is also hoped to examine the effect curvature of the metal plate in which the slot is cut.

(6) ACKNOWLEDGMENTS

We wish to thank Mr. R. G. Sinclair who ably assisted with the ttern measurements and Messrs. R. E. Walpole and W. A. pe who carried out most of the computations. Valuable lvice was also received from Dr. A. W. Adey in connection ith the admittance measurements.

(7) REFERENCES

(1) WATSON, W. H.: "Wave Guide Transmission and Antenna Systems" (Oxford University Press, 1947), Chapter 6.

(2) BOOKER, H. G.: "Slot Aerials and their Relation to Complementary Wire Aerials," *Journal I.E.E.*, 1946, **93**, Part IIIA, p. 620.

(3) STEVENSON, A. F.: "Theory of Slots in Rectangular Wave Guides," *Journal of Applied Physics*, 1948, **19**, p. 24.

(4) SCHELKUNOFF, S. A.: "Electromagnetic Waves" (Van Nostrand, 1943).

(5) CARTER, P. S.: "Circuit Relations in Radiating Systems and Applications to Antenna Problems," *Proceedings of the Institute of Radio Engineers*, 1932, **20**, p. 1004.

(6) BEGOVICH, N. A.: "Slot Radiators," *ibid.*, 1950, **38**, p. 803.

(7) WAIT, J. R.: "Field produced by an Arbitrary Slot on an Elliptic Cylinder," *Journal of Applied Physics*, 1955, **26**, p. 458.

(8) WAIT, J. R., and WALPOLE, R. E.: "Calculated Radiation Characteristics of Slots in Metal Sheets," *Canadian Journal of Technology*, 1955, **33**, p. 211.

(9) SOMMERFELD, A. N.: "Vorlesungen uber theoretische Physik," Vol. 4, Optik, Wiesbaden, Dieterich Verlag, 1950.

(10) SILVER, S.: "Microwave Antenna Theory and Design," Vol. 12, M.I.T. Radiation Laboratory Series" (McGraw-Hill, 1949).

621.317.382.029.6 : 538.632

Paper No. 1913 R
Jan. 1956

THE HALL EFFECT AND ITS APPLICATION TO POWER MEASUREMENT AT MICROWAVE FREQUENCIES

By Professor H. E. M. BARLOW, Ph.D., B.Sc.(Eng.), Member, and L. M. STEPHENSON, B.Sc.(Eng.), Student.

(*The paper was first received* 18*th May, and in revised form* 10*th June*, 1955.)

SUMMARY

The paper describes experiments carried out at 4000 Mc/s on a crystal of *n*-type germanium mounted in a resonant cavity magnetically coupled to a waveguide of rectangular section supporting an H_{01} wave mode. By this means the crystal was immersed in a comparatively strong microwave magnetic field and at the same time arrangements were made to pass through the crystal a current at right angles with that magnetic field and proportional to the electric field of the waveguide.

A Hall effect in the germanium was observed under these conditions, and the time average of the Hall e.m.f. was found, when suitable phasing adjustments were made, to bear an approximately linear relationship to the power in the waveguide, reversing its sign with reversal of direction of the power.

(1) INTRODUCTION

In previous papers[1] the application of the Hall effect in a semi-conductor to the measurement of power in an electromagnetic field was discussed, and the design of an instrument based on that principle, for operation up to 20 kc/s, was described. Preliminary work was also done in an approach to the problem of developing similar wattmeters for higher frequencies, and at 300 Mc/s a Hall effect was shown to exist in *n*-type germanium.

D'Altroy and Fan[2] have determined the conductivity and permittivity of samples of germanium at 10 000 Mc/s, but so far there appears to be no information available about Hall effect at frequencies as high as this.

The experiments described in the paper were carried out at 4 000 Mc/s on *n*-type germanium with a view to establishing the existence of the Hall effect at that frequency and to applying it, if possible, to the measurement of power in a waveguide.

(2) PRINCIPLE OF THE METHOD EMPLOYED AND CONSTRUCTION OF THE SEMI-CONDUCTOR WATTMETER

If a semi-conductor element as shown in Fig. 1 is erected in an electromagnetic field so as to produce a current i_c along the *y*-axis with a flux density B along the *z*-axis, a Hall e.m.f., v, will be set up along the *x*-axis and at any instant will be given by

$$v = \left(\frac{\mathscr{R}}{t}\right) i_c B$$

where $\mathscr{R}$ is the Hall coefficient.

When arrangements are made such that i_c is proportional to the electric component of the electromagnetic field and B is proportional to the magnetic component, the true mean value in time of v will be a measure of the power passing through the semi-conductor element. The sign of the Hall e.m.f. is dependent on the direction of the power.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

Professor Barlow is Pender Professor of Electrical Engineering at University College, London.

Mr. Stephenson is in the Department of Electrical Engineering, University College, London.

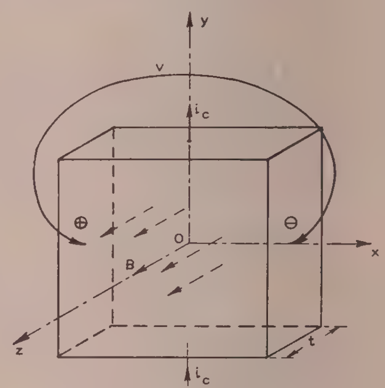


Fig. 1.—Semi-conductor element and co-ordinate system employed.

The positive and negative signs refer to the polarity of the Hall e.m.f. for mobile negative carriers.

The practicability of this device when associated with conductors forming a transmission line has already been demonstrated, and in principle there seemed no insurmountable obstacle to a similar application in a waveguide, provided that appropriate arrangements were made. The Hall effect may be regarded as arising from a sideways pressure on the mobile carriers constituting the current i_c in the semi-conductor element, and there appears to be no obvious reason why the time average of this pressure should vary widely with frequency up to microwave values.

It is, however, important to remember that with semi-conductors like the germanium used in these experiments, which had a resistivity of 37 ohm-cm and a relative permittivity of about 16, the displacement current through the material at 4 000 Mc/s is of the same order of magnitude as the conduction current. Since the skin depth at this frequency is about 0·5 cm the equivalent high-frequency circuit of the semi-conductor element is a resistance shunted by a capacitance of approximately the same impedance.

Another important aspect of the ultra-high-frequency problem is to find a method of strengthening the magnetic field applied to the crystal sufficiently to produce a measurable Hall output for powers of a few watts with only a small current i_c. As pointed out in previous papers, there is always some residual rectifier action due to asymmetry of the crystal, and this superimposes at the Hall contacts an output which, under the usual conditions of operation, is roughly proportional to the square of the potential difference across the crystal. The output from residual rectification is, of course, dependent only on the electric field component of the electromagnetic wave in which the semi-conductor element is immersed, and therefore does not reverse its sign, as the Hall effect does, when the direction of the power is reversed.

Something equivalent to the use of coils and high-permeability

naterials permissible at low frequencies for increasing the trength of the applied magnetic field has to be found for the waveguide. A simple calculation shows that, with the wave mpedances normally available, the magnetic-field component s much too small. Conditions can be improved either by deliberately setting up a standing wave and choosing the point of maximum magnetic field or by employing an E-mode not far above cut-off.

A better method is to employ a resonant cavity housing the semi-conductor unit and excited as a series element of the main rectangular H_{01} waveguide carrying the power, as shown in Figs. 2 and 3. In this arrangement the resonant cavity,

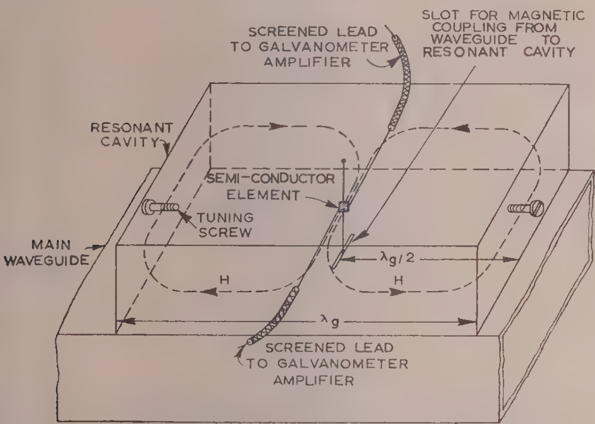


Fig. 2.—Semi-conductor element mounted in resonant cavity for microwave power measurement.

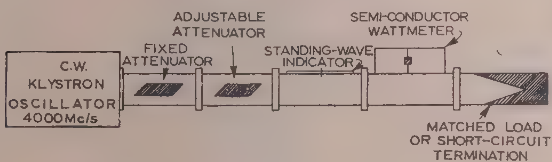


Fig. 3.—Schematic of experimental apparatus.

mounted immediately on top of the main waveguide, is about one wavelength long and is fed by a series slot at the centre of the broad face, giving almost pure magnetic coupling. The semi-conductor is erected parallel with the narrow face of the cavity and immediately opposite the slot used for excitation. In this position a strong magnetic field, built up by resonance, is applied to the semi-conductor element, but with perfect symmetry the element should not carry any current arising from the standing wave in the cavity. There is, however, a current I_0 in the semi-conductor element picked up by a wire attached to it and projecting slightly into the main waveguide through the middle of the slot in its broad face. This current is associated with a TEM wave and does not significantly excite the resonant cavity. Screws in the end walls of the resonant cavity provide for tuning and adjustment of electrical symmetry.

With this arrangement the equivalent circuits for the cavity and the associated semi-conductor element may be represented approximately as shown in Fig. 4; the vector diagrams in Fig. 5 show the phase relationships to be expected between the currents and voltages in those circuits.

The vector diagram for the cavity and the coupling slot assumes conditions approaching series resonance of the cavity by itself, but it is clear that a parallel resonance with the coupling

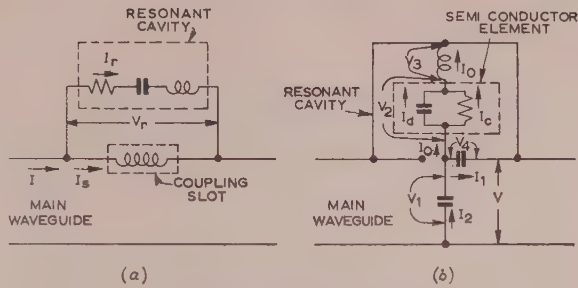


Fig. 4.—Equivalent circuits.

(a) Coupling slot and resonant cavity.
(b) Semi-conductor element and current pick-up.

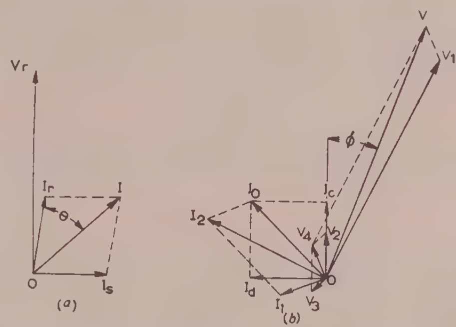


Fig. 5.—Vector diagrams for equivalent circuits.

(a) Coupling slot and resonant cavity slightly detuned from series resonance.
(b) Semi-conductor element and current pick-up.

slot can also be established. Care was taken to employ a narrow slot that was only a small fraction of a wavelength long, so that practically pure magnetic coupling between the main waveguide and the cavity was to be expected. The metal forming the wall of the slot was $\frac{1}{16}$ in thick, with an opening 1·5 cm long and 0·063 cm wide. Measurements based on the potentiometer principle[3] were made to examine the phase relationship between the fields in the cavity and in the main waveguide. Within the accuracy of $\pm 5°$ obtainable by this method the two corresponding magnetic fields were found to be in quadrature. Under the conditions of operation, the cavity current I_r in the vector diagram [Fig. 5(a)] was generally small compared with the main waveguide current I, and the phase angle θ between these currents was therefore rather less than a right angle when the cavity was adjusted to series resonance. This was borne out by experiment.

The total current I_0 through the semi-conductor element was supplied from an electric-field probe projecting not more than about 2 mm into the main waveguide, and in these circumstances one would expect the probe to be represented by a purely capacitive circuit element. There will, however, be a small series inductance associated with the pick-up wire connected to the semi-conductor element, and also a shunt capacitance between that wire and the metal wall separating the cavity from the main waveguide. The return path for the current I_0 through the wall of the cavity should be of relatively small impedance. If the conduction component, I_c, of the semi-conductor current is responsible for producing the required Hall effect, then to give the correct phasing we should expect $\theta = \phi$ in Figs. 5(a) and 5(b). Experiments to check the validity of this condition are still in progress, and no conclusive information about the

precise phase angles involved in power measurement by this method has yet been obtained. That suitable phasing can be achieved is evident from the results recorded. The Hall leads to the semi-conductor element may cause some disturbance of the conditions postulated, and it is clearly important to avoid shunt currents of any considerable magnitude passing to the wall of the cavity by way of those leads. Moreover, the electrical symmetry of the semi-conductor element is important and must be maintained by keeping the high-frequency impedances of the two Hall leads as nearly equal as possible.

It seems not altogether impossible that the displacement component, I_d, of the total semi-conductor current contributes to the Hall effect. Such a view is supported to some extent by the knowledge that radiation pressure can be observed when reflection takes place at the surface of a dielectric and also by the fact that a marked Faraday effect is exhibited in certain dielectrics. If the phase relationship of the total current in the semi-conductor and the electric field in the main waveguide can be determined precisely, it should be possible to ascertain whether the displacement component of that current does, in fact, take any part in the Hall effect.

(3) EXPERIMENTAL RESULTS

Fig. 6 shows the Hall output for different powers transmitted along the main waveguide when an *n*-type germanium crystal

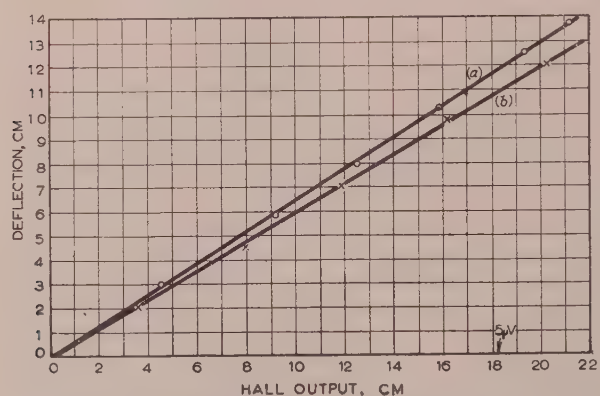


Fig. 6.—Performance of semi-conductor wattmeter.

The deflection is proportional to the power output from the electric-field probe in the main waveguide with matched load.

(*a*) Positive values of Hall output.
(*b*) Negative values of Hall output.

of resistivity 37 ohm-cm and of dimensions 0·2 cm × 0·2 cm × 0·07 cm was used in the resonant-cavity arrangement already described at a frequency of 4000 Mc/s.

The Hall output was measured by a photocell galvanometer amplifier[1] (having an input impedance of 11 kilohms with a sensitivity of 2·7 cm/μV) and figures proportional to the corresponding values of the power were deduced from the product of the minimum and maximum electric fields. No absolute values of power relating to these measurements have yet been obtained, but a rough estimate showed that a Hall output of 5 μV represented about 1 watt in the main waveguide. In these circumstances the power absorbed by the cavity was 12–15% of the power in the main waveguide and the Q-factor of the resonant cavity was thought to be about 2000.

Reversing the power through the guide reversed the Hall output, but there was a slight asymmetry, as shown in Fig. 6. This may be due to imperfections in the construction of the instrument and perhaps partly to residual rectifier action. It is hoped that further investigation will enable the asymmetry to be eliminated.

When the main waveguide was terminated in a short-circuit no measurable Hall output was obtained.

(4) CONCLUSIONS

The experiments described establish without doubt the existence of a Hall effect in *n*-type germanium when immersed in an electromagnetic field at 4000 Mc/s, but the magnitude of the corresponding Hall coefficient under these conditions has not yet been ascertained.

A method has been devised by which the high-frequency Hall effect is used to measure power transmitted along a waveguide, and a preliminary design of instrument for this purpose has been made. Further work is required to elucidate some of the conditions of operation and to develop the measuring technique. It is not yet known what degree of accuracy is obtainable in the use of this instrument, or what bandwidth it will cover. There can be no doubt, however, that a new method of power measurement has been established and that it offers the great advantage that it can be used under any conditions of load, matched or otherwise, without absorbing more than a small fraction of the power being transmitted.

(5) ACKNOWLEDGMENTS

The authors are indebted to the Trustees of the Paul Instrument Fund for financial support of this work. They also wish to acknowledge their gratitude to Professor A. L. Cullen, Dr. R. A. Smith, Dr. G. G. Macfarlane and Mr. H. G. Effemey for many helpful suggestions and discussions during the course of the investigation.

(6) REFERENCES

(1) BARLOW, H. E. M.: "The Application of the Hall Effect in a Semi-Conductor to the Measurement of Power in an Electromagnetic Field" and "The Design of Semi-Conductor Wattmeters for Power-Frequency and Audio-Frequency Applications," *Proceedings I.E.E.*, Papers No. 1654 M, June, 1954, and 1778 M, November, 1954 (**102 B**, pp. 179 and 186).

(2) D'ALTROY, F. A., and FAN, H. Y.: "Microwave Measurements on Germanium Semi-Conductors," *Proceedings of the National Electronics Conference* (Chicago), 1952, **8**, p. 522.

(3) GOODENOUGH, E. F.: Discussion on "Radio Measurements in the Decimetre and Centimetre Wavebands," *Journal I.E.E.*, 1946, **93**, Part III, p. 117.

21.396.93

Paper No. 1960 R
Jan. 1956

THE THEORETICAL DESIGN OF DIRECTION-FINDING SYSTEMS FOR HIGH FREQUENCIES

By W. C. BAIN, M.A., B.Sc., Ph.D.

(*The paper was first received* 14*th July, and in revised form* 24*th September*, 1955.)

SUMMARY

The problem of finding the bearing of a distant h.f. transmitter in onditions of wave interference is examined for the simplified case of on-interacting aerials on a plane earth and with no pick-up of horiontally polarized radiation. Two methods of approach are considered –solution of the field equations for a number of incident plane waves rom a knowledge of the field at the aerials, and the fitting of rectilinear onstant-phase lines to observed values by a "least squares" process. t is shown that the cyclical system of Earp and Godfrey is a "least quares" system of the latter type. Systems of the Wullenwever ind bear a close resemblance to a least-squares system with weighting according to the signal amplitude at each aerial; the difference lies n the fact that they operate with sinusoidal functions of phase instead f linear functions.

(1) INTRODUCTION

The use of a wide-aperture aerial system appears to give the rospect of a high-frequency direction-finder with much better erformance than the present narrow-aperture instruments. This ossibility has been realized for a number of years—see, for nstance, Ross[1]—and a start has been made in constructing and esting such systems. There is, for example, the direction-finder mploying cyclical measurement of phase described by Earp and Godfrey;[2] some trials on a system of this type are reported by Hopkins and Bramley.[3] Still earlier, the German instrument nown as Wullenwever[4] was in operation during the 1939–45 Var. It is therefore of interest to make a more general approach o the problem of designing a wide-aperture system.

Now this problem can be divided into three main parts. The nitial assumption can be made that on an ideal plane site aerials re available which pick up only one component of the field of he incident ionospheric wave—say the vertically polarized comonent—and which do not interact appreciably. The first part f the problem consists in devising a method of utilizing these ignals in such a way as to obtain the required bearing information. In general, the incident wave will contain several ionopheric modes, and ideally the direction-finder should give the earings of each of these independently. However, for most urposes the bearing of the strongest mode or the mean bearing f all the modes (weighted according to signal strength) would be lmost as satisfactory, because the strongest mode will usually be hat which has travelled from transmitter to direction-finder with he smallest possible number of hops, and will therefore be the east laterally deviated. The weighted mean bearing of the modes vill probably be quite close to the bearing of the strongest mode; may be a little better or worse than the latter as an indication f the true transmitter bearing according to the relative magnitude f the standard deviations of the bearing deviations in each node. The paper is chiefly concerned with ways of solving this art of the problem.

Next comes the devising of a circuit technique which will andle the incoming signals in the required manner. Although probably any proposed scheme could be made to function by electronic means, in many cases the resulting equipment would be too expensive to be worth serious consideration.

The third part of the problem consists in overcoming the difficulties which arise through substituting real aerials and feeders and an actual site in place of the ideal situation previously considered. This may lead to errors arising through aerial interaction, reception of the undesired component of polarization, or from interference by atmospheric noise or by unwanted transmitters. Re-radiation from surrounding objects may also give errors, although one would normally expect that the measures taken in the design of the aerial system to deal with the intermode wave-interference problem would be adequate to make the site errors negligible.

(2) SYSTEMS IN WHICH THE EQUATIONS OF THE WAVE INTERFERENCE FIELD ARE SOLVED

Methods of deriving bearing information from ideal aerials placed in the incident field will now be considered. The most obvious d.f. system which will operate in a region of wave interference is perhaps the following: Assume that there are N different ionospheric plane waves incident on the system, where N is a small number. If one of these waves is taken as a reference in phase and amplitude, the system is specified completely by $(4N - 2)$ parameters; these may be taken to be the bearing and elevation of the reference wave, and the phase, amplitude, bearing and elevation of the others at a fixed point. If $(n + 1)$ aerials are then placed in the field, including a reference aerial, one can obtain $2n$ measurements of phase and amplitude relative to the reference aerial. Each of these measurements can be related by an equation to the $(4N - 2)$ parameters of the incident waves. If the number of equations is equal to the number of parameters it seems that, in general, the $(4N - 2)$ parameters could be found. In this case the relation between n and N is given by

$$2n = 4N - 2$$

i.e.
$$n + 1 = 2N \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

Thus if there are two incident waves, enough information should be obtainable from four aerials to specify their bearings, elevations and their relative phase and amplitude.

However, the solution of the equations is not as simple as may appear from the above account. In the first place, a unique solution may not be possible because the phase can be measured only within a range of 2π radians. Another type of difficulty is illustrated by the case of a single incident plane wave. Eqn. (1) indicates that two aerials should suffice to give its bearing and elevation, but this is not so; the trouble here lies in the fact that the amplitude measurement provides no useful information, since the amplitude is constant everywhere. Again, difficulties may arise because each mode of propagation does not consist of a plane wave but of plane waves whose directions of travel are contained within a cone of small angle. However, even for a small number of incident plane waves the analytical difficulties in the problem are formidable.

Written contributions on papers published without being read at meetings are ivited for consideration with a view to publication.
The paper is an official communication from the Radio Research Station, Departent of Scientific and Industrial Research.

Now that it has been pointed out that a solution of the d.f. problem may be found in this way, although the prospects are not encouraging, another line of approach will be examined.

(3) SYSTEMS IN WHICH LINES OF CONSTANT PHASE ARE FITTED TO THE OBSERVED FIELD

(3.1) General Considerations

The field of a single plane wave at the surface of the ground contains a number of equally spaced straight lines of constant phase, which are normal to the direction of propagation. Even if other waves are introduced, provided that their combined amplitude does not exceed that of the main wave, the constant-phase lines, although distorted somewhat in shape, still oscillate about the position of the original phase lines (see Section 7.1). If there is no single wave in the incident radiation which is larger in amplitude than all the others combined, the situation is not so simple. However, it can still be said that over the ground the phase of the resultant signal will resemble most closely that of the strongest wave.

It therefore seems that a good value for the bearing of the incident waves might be obtained by attempting to fit a set of equally spaced straight constant-phase lines to observed values of phase at various aerials, and this will now be tried; i.e. an attempt will be made to find the "best" plane wave from the given data. In problems involving the fitting of curves to data, least squares methods are commonly applied and will be used here.

It may be asked whether such a system would give the bearing of the strongest mode or of some particular combination of the modes. In wave-interference conditions it must be expected that the indicated bearing will vary with the position of the system on the ground. One cannot therefore answer the question, but can only ask in its place whether the mean indicated bearing has a particular value, or the bearing as the system aperture tends to infinity. The latter concept is simpler to use, for if one ray is stronger than all the others the indicated bearing will tend to that of the strongest mode as the system aperture tends to infinity. When no one ray is greater in amplitude than all the others, it is not clear to what value the indicated bearing will tend. Of course, the bearing error might still be taken to be the difference between the indicated bearing and the bearing of the strongest mode, but a fairer method to establish an average bearing error for the system would be to proceed as follows:

Consider a case where two rays are incident on a direction-finder with relative amplitude r and angles of elevation δ_1 and δ_2. Suppose that the true bearing of the transmitter is θ, that the bearings of the two incident rays are $(\theta + \alpha_1)$ and $(\theta + \alpha_2)$, and that the phase difference of the two rays at a fixed point is ϕ. Let the indicated bearing of the system be γ. Now in practical direction-finding one is concerned with the departure of the indicated bearing from the true transmitter bearing, so a good measure of the error of the direction-finder is the average value of $(\gamma - \theta)^2$. This will have to be averaged over the values of α_1, α_2, and ϕ which are found in practice. For a theoretical calculation, it seems reasonable to give α_1 and α_2 normal distributions with zero mean and standard deviations σ_1 and σ_2, and to give ϕ a uniform distribution from 0 to 2π. The resulting figure will be the error variance of the system for the particular values of r, δ_1 and δ_2 concerned. In this way the performance of a system may be assessed, even if in wave-interference conditions it is difficult to see whether it is attempting to give the bearing of the strongest mode or of some combination of modes. The method can readily be extended in principle to deal with more than two incident rays. For the present, however, it will simply be assumed that the least-squares process will give a low error variance in the above sense.

Suppose there are n aerials in addition to a reference aerial. Let the amplitude and phase of the total signal at the kth aerial be V_k and ψ_k relative to that at the reference aerial; the kth aerial will be taken to be situated at the point (x_k, y_k) in Cartesian co-ordinates, the reference aerial being at the origin. Bearings will be measured here for convenience anti-clockwise from the x-axis, to conform with the usual mathematical practice. If it is desired to think in terms of bearings measured clockwise from north, the formulae given in the paper still apply if one considers the x-axis to point north and the y-axis to point east, i.e. if a left-handed pair of axes is used; θ_0 is then the bearing in the usual d.f. sense. Now, if a wave is incident on the system with bearing θ_0 and elevation δ_0, and it has a phase χ_0 at the origin, its phase, ϕ_k, at the kth aerial will be given by

$$\phi_k = \frac{2\pi}{\lambda} \cos \delta_0 (x_k \cos \theta_0 + y_k \sin \theta_0) + \chi_0 \quad . \quad (2)$$

where λ is the wavelength of the radio waves.

Phase advances are taken as positive.

(3.2) Linear Functions of Phase

The sum of the squares of the deviations of each of the observed phases from the value predicted by eqn. (2) is given by

$$S_1 = \sum_{k=1}^{n} [K_0(x_k \cos \theta_0 + y_k \sin \theta_0) + \chi_0 - \psi_k]^2 \quad . \quad . \quad (3)$$

where $$K_0 = \frac{2\pi}{\lambda} \cos \delta_0$$

The bearing indicated by the least-squares method is the value of θ_0 which makes S_1 a minimum, S_1 being also minimized with respect to K_0 and χ_0.

The process of fitting straight constant-phase lines to observed values is illustrated in Fig. 1. Four aerials and a reference aerial are shown, and for simplicity they are spaced apart by distances much less than a wavelength. Fig. 1(*a*) shows the constant-phase lines of an electromagnetic field separated by intervals of 20° of phase, and Fig. 1(*b*) the straight constant-phase lines to be fitted to the observed values at the four aerials. Fig. 1(*c*) gives the result of the fitting process; comparison of the phases indicated in brackets at each aerial in Figs. 1(*a*) and 1(*c*) reveals the closeness of fit obtained in this particular case.

Unfortunately this system does not always function if the measured value in the range $-\pi$ to π is used for ψ_k. The quantity ϕ_k is not in general limited to this range and there is no possibility of matching ϕ_k and ψ_k unless the appropriate number of 2π's can be added to the measured phase in obtaining ψ_k. One way of doing this is to place the aerials so that the distance between each aerial and its nearest neighbour is well below $\lambda/2$. Then even under wave-interference conditions the phase difference between the signals at neighbouring aerials will rarely exceed π radians; and starting from the reference aerial, one can find the phase at any aerial relative to it by adding multiples of 2π so as to keep every phase difference between neighbouring aerials below π in magnitude.

A system of this type with close spacing of aerials might function in either of two ways. S_1 in eqn. (3) might be generated electronically and the minimum found by some searching process involving variation of the parameters θ_0, K_0 and χ_0. An alternative method is to solve the equations

$$\frac{\partial S_1}{\partial \theta_0} = \frac{\partial S_1}{\partial K_0} = \frac{\partial S_1}{\partial \chi_0} = 0$$

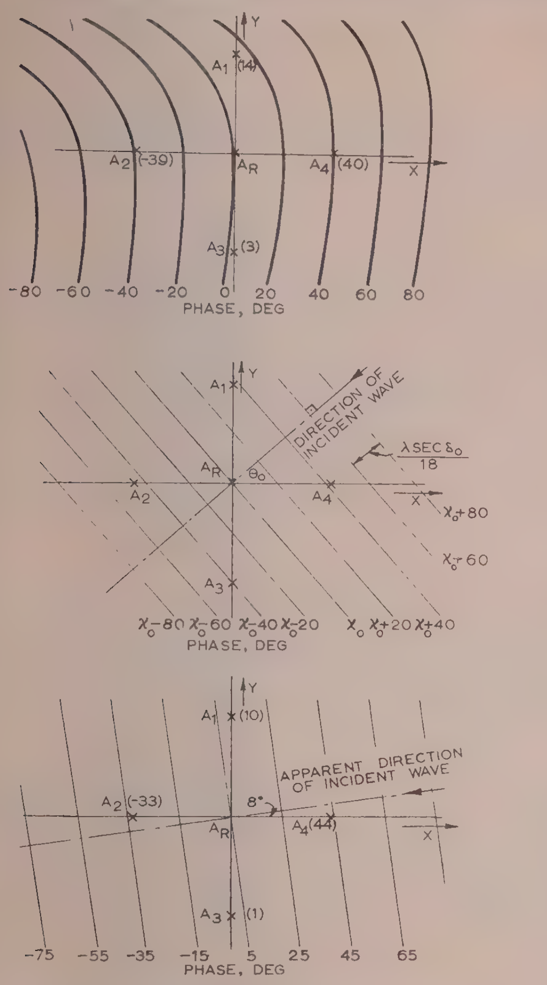


Fig. 1.—To illustrate the process of fitting straight constant-phase lines to observed phases at four aerials, A_1, A_2, A_3 and A_4; A_R is a reference aerial at the origin.

(*a*) A portion of a family of constant-phase lines. The phases which would be measured at each aerial relative to A_R are given in brackets.
(*b*) A set of straight constant-phase lines which are to be fitted to the observed phases at each aerial. δ_0 has been taken to be 0° here.
(*c*) The straight constant-phase lines fitted to the observed values in (*a*) by minimizing the sum S_1. The phases given in brackets beside the aerial positions are the values obtained by interpolation from these lines.

hereby obtaining an explicit relation for θ_0 in terms of the measured phases, and to construct circuits which will provide the value of θ_0 from this relation. This type of system will be considered in a little more detail. The value of $\partial S_1/\partial\theta_0$ is given by

$$\begin{aligned}\frac{\partial S_1}{\partial\theta_0} &= \sum_k 2[K_0(-x_k\sin\theta_0 + y_k\cos\theta_0)] \\ &\qquad [K_0(x_k\cos\theta_0 + y_k\sin\theta_0) + X_0 - \psi_k] \\ &= 2K_0[-\tfrac{1}{2}K_0\sin 2\theta_0(\Sigma x_k^2 - \Sigma y_k^2) + K_0\cos 2\theta_0\Sigma x_k y_k \\ &\quad - X_0\sin\theta_0\Sigma x_k + \sin\theta_0\Sigma x_k\psi_k + X_0\cos\theta_0\Sigma y_k - \cos\theta_0\Sigma y_k\psi_k] \\ &= 0\end{aligned}$$

when S_1 is a minimum.

If the aerials are symmetrically disposed about the x- and y-axes, say in a circle with centre at the origin,

$$\Sigma x_k^2 = \Sigma y_k^2$$

and

$$\Sigma x_k y_k = \Sigma x_k = \Sigma y_k = 0$$

If the aerials are in a circle of radius greater than $\lambda/2$ it will be supposed for the moment that other aerials are placed along a radius of the circle spaced less than $\lambda/2$ apart so as to keep track of the cycles of phase. The phases measured at these aerials are not used in the summations. Hence

$$\frac{\partial S_1}{\partial\theta_0} \doteq 2K_0(\sin\theta_0\Sigma x_k\psi_k - \cos\theta_0\Sigma y_k\psi_k)$$

Equating this to zero gives

$$\tan\theta_0 = \frac{\Sigma y_k\psi_k}{\Sigma x_k\psi_k}$$

Since this is independent of K_0 and X_0 there is no need to consider the other equations in $\partial S_1/\partial K_0$ and $\partial S_1/\partial X_0$.

If the positions of the aerials in the ring are given by

$$x_k = a\cos\frac{2k\pi}{n} \quad \text{and} \quad y_k = a\sin\frac{2k\pi}{n}$$

$$\tan\theta_0 = \frac{\Sigma\psi_k\sin\dfrac{2k\pi}{n}}{\Sigma\psi_k\cos\dfrac{2k\pi}{n}}$$

it can easily be shown that the true minimum value of S_1 is given by the relation

$$\theta_0 = \arg\left[\sum_{k=1}^{n}\psi_k\exp\left(\frac{2k\pi j}{n}\right)\right] \quad . \quad . \quad . \quad (4)$$

which resolves an ambiguity of π in the value of θ_0 given by the $\tan\theta_0$ expression.

A method of solving this equation by electronic means is known, for in Section 7.2 it is shown that the bearing indicated by the cyclical differential system of Earp and Godfrey[2] is identical with eqn. (4). It is therefore true that this particular cyclical system is one which attempts to fit equally-spaced straight constant-phase lines to the observed phases by a least-squares process.

It is interesting to note that the bearing in the cyclical system is actually produced by operations only with phase differences between aerials in the ring, so that the reference aerial and linking aerials are unnecessary in this case. The fact that only phase differences are involved can be seen from eqn. (18) in Section 7.2, this being equivalent to eqn. (4).

In the simplest form of the Earp system the aerials are spaced by less than $\lambda/2$, so that it is strictly of the type considered here. However, to cover higher frequencies with the same aerials Earp and Godfrey propose a system of "phase compression" which will function with aerial spacings somewhat greater than $\lambda/2$, but which also has an indicated bearing given by eqn. (4). This brings out the point that there are methods of resolving the phase ambiguities other than the use of aerials spaced by less than $\lambda/2$. One might, for instance, use a pair of aerials spaced $0{\cdot}4\lambda$ apart and another pair in the same direction spaced 4λ apart. If the phase difference measured at the former pair were ψ, we might predict a phase difference of 10ψ at the latter pair; one could then add multiples of 2π to the measured phase difference at the widely spaced pair to bring the value as near to 10ψ as possible. The "phase compression" scheme itself illustrates another way of overcoming the difficulty, and there may well be still more of a similar type.

All the methods of overcoming the phase-ambiguity problem can fail under conditions of severe wave interference. Even if the aerial spacing is well below $\lambda/2$, the phase difference between aerials will certainly exceed π sometimes, and quite serious errors may be introduced. If the aerial spacing exceeds $\lambda/2$, it seems likely that incorrect prediction of the number of phase cycles will

often occur unless an ample margin of safety is allowed in the predicting device.

The possibility of weighting the terms in the sum of squares according to the amplitudes of the signals at each aerial should not be overlooked. The most distorted parts of the phase field are undoubtedly those where the amplitude of the signal is smallest. If the phases in these parts of the field are made to have little weight in the sum of squares, the straight lines which are fitted to them will lie nearer to the constant-phase lines in the least distorted regions, where they will be more nearly normal to the bearing of the strongest mode. It would therefore probably be advantageous to minimize a sum such as

$$S_2 = \sum_{k=1}^{n} E_k[K_0(x_k \cos\theta_0 + y_k \sin\theta_0) + \chi_0 - \psi_k]^2 \quad . \quad (5)$$

but this particular weighted sum will not be considered further here.

(3.3) Sinusoidal Functions of Phase

Every system considered in Section 3.2 contains some device which in effect predicts the number of cycles of phase involved in going from one aerial to the others. However, even if no attempt is made to do this, it is still possible to apply a least-squares method. One might, for example, minimize the sum

$$S_3 = \sum_{k=1}^{n} (\phi_k - \psi_k + 2\pi m_k)^2 \quad . \quad . \quad . \quad (6)$$

where m_k is an integer chosen to satisfy

$$|\phi_k - \psi_k + 2\pi m_k| \leqslant \pi \quad . \quad . \quad . \quad . \quad (7)$$

ϕ_k is given by eqn. (2). When θ_0 is in the neighbourhood of the true bearing, S_3 is likely to be the same as S_1 and therefore to give the same indicated bearing.

As θ_0 is varied in this system the values of m_k will have to undergo a series of discontinuous jumps to satisfy eqn. (7), and, in general, S_3 will exhibit a number of minima. The lowest-valued of these should be taken as giving the indicated bearing. For a single incident plane wave with bearing θ, S_3 would be zero for $\theta_0 = \theta$; it might be zero for other values of θ, but with enough aerials present it is likely that the ambiguous values could be eliminated.

Eqn. (6) could no doubt be set up and minimized by electronic means, despite the presence of the awkward quantities m_k. However, there is a simpler equation which can be considered in its place and which is very similar to it, at least for small departures of ψ_k from a plane-wave state. This is

$$S_4 = \sum_{k=1}^{n} [\sin(\phi_k - \psi_k)]^2 \quad . \quad . \quad . \quad . \quad (8)$$

The essential point about the function $(\phi_k - \psi_k + 2\pi m_k)$ in S_3 was that it was periodic in 2π; the use of a sine function also gives this periodicity without the necessity of introducing m_k. The quantity $\sin(\phi_k - \psi_k)$ falls to zero for $|\phi_k - \psi_k| = \pi$ as well as 0, and it is therefore advantageous to use an expression such as $\left[2\sin\left(\frac{\phi_k - \psi_k}{2}\right)\right]^2$ in the sum; this tends to $(\phi_k - \psi_k)^2$ when $(\phi_k - \psi_k)$ is small, as does $\sin^2(\phi_k - \psi_k)$, but does not become zero again until $|\phi_k - \psi_k|$ reaches 2π. It is a single-valued function of $\cos(\phi_k - \psi_k)$ and so can be measured. It therefore appears that a more useful sum to minimize is given by

$$S_5 = \sum_{k=1}^{n} 4\sin^2\left(\frac{\phi_k - \psi_k}{2}\right),$$
$$= 2\Sigma[1 - \cos(\phi_k - \psi_k)] \quad . \quad . \quad . \quad . \quad (9)$$

Clearly one would obtain the same result by maximizing

$$S_6 = \sum_k \cos(\phi_k - \psi_k) \quad . \quad . \quad . \quad . \quad (10)$$

Now consider the effect of amplitude weighting in this case; this requires the introduction of a factor E_k into each term in eqn. (9). Mimimizing such a sum is the same as maximizing the expression

$$S_7 = \sum_k E_k \cos(\phi_k - \psi_k) \quad . \quad . \quad . \quad . \quad (11)$$

Such a system bears a marked resemblance to a direction-finder which has already been constructed—the Wullenwever.[4] This has a ring of 40 aerials, and in one mode of operation 12 of these are combined to give direction-finding by means of rotating the maximum of the aerial polar diagram. The radio waves arriving with phase ψ_k at the kth aerial are phase-shifted by ϕ_k, and the indicated bearing is taken to be where the sum

$$\Sigma E_k \cos(\omega t + \psi_k - \phi_k)$$

has maximum amplitude. This amplitude may be written

$$S_8 = \left|\sum_k E_k \varepsilon^{j(\psi_k - \phi_k)}\right| \quad . \quad . \quad . \quad . \quad . \quad (12)$$

S_8 is obviously closely related to S_7; the exact connection will be stated shortly. There is, however, no provision for varying δ_0 in the Wullenwever, so it rather falls short of the ideal least-squares system. The other method of direction-finding in this system, in which two groups of six aerials are used to give a minimum characteristic, will not be considered here.

Another form of the d.f. system which consists in maximizing S_8, but including the variation of δ_0 as well as θ_0, is worthy of further consideration, for published data are available which help in assessing its performance. It employs phase-shifting of the signals received at every aerial in a ring of equally spaced aerials, n in number. Let a plane wave, with elevation δ and bearing θ, be incident on the system, which has diameter d, so that

$$\psi_k = \frac{\pi d}{\lambda}\cos\delta\cos\left(\theta - \frac{2k\pi}{n}\right) \quad . \quad . \quad . \quad (13)$$

Now $\phi_k = \frac{\pi d}{\lambda}\cos\delta_0\cos\left(\theta_0 - \frac{2k\pi}{n}\right)$, from eqn. (2) omitting the term χ_0, which does not alter the value of S_8 given in eqn. (12). The aerial polar diagram of such a system has been considered by Knudsen[5]; the following parameters are useful in studying it. S_8 is, of course, proportional to the polar diagram function.

$$\beta = \frac{\pi d}{\lambda}$$

$$t^2 = \cos^2\delta + \cos^2\delta_0 - 2\cos\delta\cos\delta_0\cos(\theta - \theta_0)$$

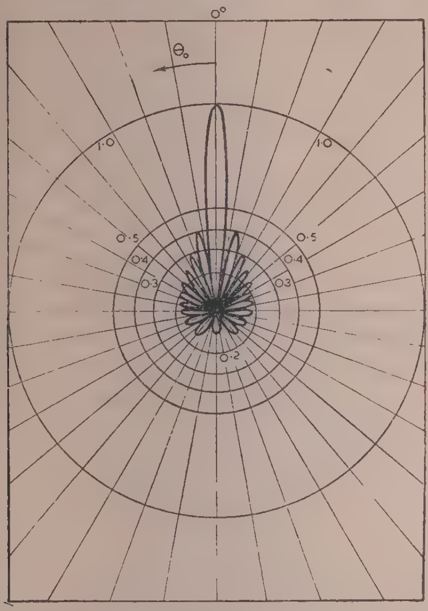
[Note: $t = 2\sin\left(\frac{\theta - \theta_0}{2}\right)$ if $\delta = \delta_0 = 0$.]

$$\tan\xi = \frac{\cos\delta\sin\theta - \cos\delta_0\sin\theta_0}{\cos\delta\cos\theta - \cos\delta_0\cos\theta_0}$$

For an infinite number of aerials S_8 is given by

$$S_8 = \text{Constant} \times J_0(\beta t) \quad . \quad . \quad . \quad . \quad (14)$$

A plot of this function against θ_0 in a particular case is shown in Fig. 2; the shape of the diagram for varying θ is the same. The largest maximum is in the direction $\theta = \theta_0$ and there are subsidiary maxima falling off in amplitude as $|\theta_0 - \theta|$ is increased.



. 2.—The amplitude polar diagram of the S_8 system for parameter values $\delta = \delta_0 = 0$, $\theta = 0$, $d/\lambda = 5$, and $n = \infty$.

f a finite odd number n of aerials is used, the polar diagram of array is given by

$$\beta t) + 2\sum_{k=1}^{n} J_{2kn}(\beta t) \cos\left[\left(\frac{\pi}{2} - \xi\right)2kn\right]$$
$$- j2\sum_{k=0}^{\infty} J_{(2k+1)n}(\beta t) \sin\left[\left(\frac{\pi}{2} - \xi\right)(2k+1)n\right] \quad . \quad (15)$$

The first-order correction term is in the second summation only tting $k = 0$) and is in $J_n(\beta t)$. This is negligible if $\beta t < n$, and may always be neglected if

$$\frac{2\pi d}{\lambda} < n$$

is implies that the aerial spacing should be less than $\lambda/2$—a ndition which often occurs in these investigations. In other rds, if the aerial spacing is below $\lambda/2$, the polar diagram is actically the same as that for an infinite number of aerials. hen the spacing is increased above this, subsidiary lobes of preciable amplitude appear at $|\theta_0 - \theta| = \pi$ and move round vards $|\theta - \theta_0| = 0$.

This system appears to be sufficiently promising to merit nsideration as a practical direction-finder; indeed a system of S_8 type with a rather more elaborate aerial array has already en proposed by Bray.[6] Many other problems arise in devising oractical system, but these will not be considered here. It will ly be pointed out that, since it represents a close approach to a st-squares system with amplitude weighting, there is some son for believing it to be capable of a better performance than cyclical system as proposed by Earp and Godfrey,[2] which is east-squares system without amplitude weighting.

It should also be noted that the system obtained by maximizing has certain advantages over the S_8 system. For

$$S_7 = \Sigma E_k \cos(\psi_k - \phi_k)$$
$$= J_0(\beta t) + 2\sum_{k=1}^{\infty} J_{2kn}(\beta t)\cos\left[\left(\frac{\pi}{2} - \xi\right)2kn\right] \quad . \quad (16)$$

if the aerials lie in a circle. The first-order correction term here is in $J_{2n}(\beta t)$ instead of $J_n(\beta t)$ as with the S_8 system. The S_7 system will therefore give better results than the S_8 system with the same number of aerials, or similar results with fewer aerials. These remarks apply to the side-lobes only; for aerial spacings less than λ, say, the indicated bearings will be the same on both the S_7 and S_8 systems.

It can be seen from the above that any direction-finding system operating by rotation of a polar diagram with a narrow main lobe must be closely related to a least-squares system with amplitude weighting. If such a system could be made with a main lobe occupying less than one degree in azimuth and with no appreciable side-lobes, the individual incident rays could be resolved; they would appear as distinct maxima in S_8 as θ_0 and δ_0 were varied. Such a polar diagram could probably not be realized in the h.f. band, but systems falling far short of this ideal are still of value in the problem of locating a transmitter.

(4) CONCLUSIONS

Two very broad methods of approach to the problem of finding the bearing of a distant h.f. transmitter are considered; the first is to deduce from observations at a few aerials the bearing and elevation of all incident plane waves, and the second to attempt to fit straight constant-phase lines by a least-squares method to observed values of phase at a number of points. The latter method appears to be easier to handle. In certain cases it tends to give the bearing of the strongest incident mode, but when there are several incident waves of comparable amplitude it does not tend to a definite bearing. The indicated bearing is probably still valuable, however; a method of assessing the error variance of a system about the true transmitter bearing is outlined in the paper.

It has been shown that the cyclical system of Earp and Godfrey in effect attempts to fit straight constant-phase lines to observed phases by a least-squares process. Systems of the Wullenwever type are least-squares systems with amplitude weighting, although they use sinusoidal functions of phase instead of linear ones. The amplitude weighting may give such systems an advantage over the Earp system, although they have not been constructed in a form which will give a mathematically explicit value for the bearings, which the Earp system does produce.

(5) ACKNOWLEDGMENTS

The work described in the paper was carried out as part of the programme of the Radio Research Board. The paper is published by permission of the Director of Radio Research of the Department of Scientific and Industrial Research.

(6) REFERENCES

(1) Ross, W.: "Fundamental Problems in Radio Direction Finding at High Frequencies," *Journal I.E.E.*, 1947, **94**, Part IIIA, p. 154.
(2) Earp, C. W., and Godfrey, R. M.: "Radio Direction-Finding by the Cyclical Differential Measurement of Phase," *ibid.*, 1947, **94**, Part IIIA, p. 705.
(3) Hopkins, H. G., and Bramley, E. N.: "Some Practical Measurements of the Relative Performances of a Cyclical Phase-Comparison Type of Direction-Finder and a U-Adcock Instrument," *Proceedings I.E.E.*, Paper No. 1544 R, September, 1953 (**100**, Part III, p. 263).
(4) "Radio Direction Finding and Navigational Aids," D.S.I.R. Radio Research Special Report No. 21 (H.M. Stationery Office, 1949).

(5) KNUDSEN, H. L.: "The Necessary Number of Elements in a Directional Ring Aerial," *Journal of Applied Physics*, 1951, **22**, p. 1299.

(6) BRAY, W. J.: British Patent No. 696281. 1949.

(7) APPENDICES

(7.1) A Theorem concerning Constant-Phase Lines

It is here shown that constant-phase lines on the ground depart by less than $(\lambda/4)\sec\delta$ from the corresponding lines for their strongest component plane wave, provided that this wave is of greater amplitude than all the others combined in phase.

Suppose that the strongest plane wave has zero phase at a reference point R in the field; and suppose for the moment that only this wave is present. Consider any constant-phase line L_1 on the ground, say with phase ψ relative to R. It will have on either side of it two other parallel constant-phase lines L_2 and L_3, with phases $(\psi + \pi/2)$ and $(\psi - \pi/2)$ respectively; these are straight lines, as is L_1, and are separated from L_1 by a distance $(\lambda/4)\sec\delta$.

Suppose now that other waves are introduced into the field, the sum of whose amplitudes is less than that of the original. On L_2 the phase of the resultant can never change by as much as $\pi/2$ from its original value of $(\psi + \pi/2)$ and so at no point on L_2 does the phase reach ψ. This means that the new constant-phase line with phase ψ cannot cross or even touch L_2. Similarly it cannot cross or touch L_3. Hence the constant-phase line with phase ψ cannot depart from its original position (which is L_1) by more than $(\lambda/4)\sec\delta$, and the theorem is proved.

(7.2) The Indicated Bearing on the Cyclical System of Earp and Godfrey

In the basic regime of this system the phase difference between successive aerials is measured in turn round the circle, the cycle being performed at an angular frequency ω_1. Suppose the phase at the kth aerial is ψ_k. Then if a linear phase discriminator is used, a voltage is produced which consists of a succession of values proportional to $(\psi_k - \psi_{k+1})$. Let this voltage be represented by $f(x)$, where $x = \omega_1 t$, t being a time variable. Then, ignoring the constant of proportionality, which does not affect the argument, we have

$$f(x) = \psi_k - \psi_{k+1}, \text{ where } \frac{2k\pi}{n} \leqslant x < \frac{2(k+1)\pi}{n}$$

and $$k = 0, 1, 2 \ldots (n-1)$$

Now expand $f(x)$ as a Fourier series, as the bearing is obtained from the phase of the fundamental in $f(x)$ in this system.

$$f(x) = a_1 \cos x + b_1 \sin x + a_2 \cos 2x + \ldots$$

Then $$\pi a_1 = \int_0^{2\pi} f(x)\cos x\,dx$$

$$\pi b_1 = \int_0^{2\pi} f(x)\sin x\,dx$$

If only one ray is incident on the system then

$$\psi_k = \beta\cos\left(\theta - \frac{2k\pi}{n}\right)$$

where $$\beta = \frac{\pi d}{\lambda}\cos\delta$$

Hence

$$\pi a_1 = -2\beta\sin\frac{\pi}{n}\sum_{k=0}^{n-1}\int_{\frac{2k\pi}{n}}^{\frac{2(k+1)\pi}{n}} \sin\left[\theta - \frac{(2k+1)\pi}{n}\right]\cos x\,dx$$

$$= -2\beta_1\sin^2\left(\frac{\pi}{n}\right)\left\{\sum_k \sin\theta + \sum_k \sin\left[\theta - \frac{2(2k+1)\pi}{n}\right]\right\}$$

$$= -2n\beta_1\sin^2\left(\frac{\pi}{n}\right)\sin\theta$$

as the second sum vanishes for $n > 2$.

Similarly,

$$\pi b_1 = 2n\beta_1\sin^2\left(\frac{\pi}{n}\right)\cos\theta$$

If we let the indicated bearing γ be given by

$$\gamma = \arc\tan\left(\frac{-a_1}{b_1}\right) \quad . \quad . \quad . \quad . \quad . \quad (1$$

γ will assume the correct value θ when only one ray is inciden on the system. Note that the 180° ambiguity in this expressio for γ can be resolved by taking $\sin\gamma$ to have the sign of th numerator in eqn. (17), and $\cos\gamma$ the sign of the denominato γ can clearly be derived by a phase measurement upon $f(x)$.

Now suppose ψ_k is an observed value in a general inciden field. Again we have

$$\pi a_1 = \int_0^{2\pi} f(x)\cos x\,dx$$

$$= \sum_k \int_{\frac{2k\pi}{n}}^{\frac{2(k+1)\pi}{n}} (\psi_k - \psi_{k+1})\cos x\,dx$$

$$= 2\sin\frac{\pi}{n}\sum_k(\psi_k - \psi_{k+1})\cos\frac{(2k+1)\pi}{n}$$

Similarly, $$\pi b_1 = 2\sin\frac{\pi}{n}\sum_k(\psi_k - \psi_{k+1})\sin\frac{(2k+1)\pi}{n}$$

Hence the indicated bearing under these conditions is

$$\gamma = \arc\tan\left[\frac{-\Sigma(\psi_k - \psi_{k+1})\cos\frac{(2k+1)\pi}{n}}{\Sigma(\psi_k - \psi_{k+1})\sin\frac{(2k+1)\pi}{n}}\right]$$

with the 180° ambiguity resolved as above.

This can also be written

$$\gamma = \arg\left\{\sum_k(\psi_k - \psi_{k+1})\exp j\left[\frac{(2k+1)\pi}{n} - \frac{\pi}{2}\right]\right\} \quad (18$$

a method of expression which is unambiguous. Eqn. (18) ca be simplified still further:

$$\gamma = \arg\left\{\exp\left(-\frac{j\pi}{2}\right)\sum_{k=0}^{n-1}\right.$$

$$\left.\left[\psi_k\exp\frac{j(2k+1)\pi}{n} - \psi_k\exp\frac{j(2k-1)\pi}{n}\right]\right.$$

ce the sum of n successive terms in k is the same whatever value k is the first.

$$\gamma = \arg\left[(-j)\sum_k \psi_k \exp\left(\frac{2k\pi j}{n}\right)(\varepsilon^{\frac{j\pi}{n}} - \varepsilon^{\frac{-j\pi}{n}})\right]$$

$$= \arg\left[(-j)\sum_k \psi_k \exp\left(\frac{2k\pi j}{n}\right)2j\sin\frac{\pi}{n}\right]$$

$$= \arg\left[\sum_k \psi_k \exp\left(\frac{2k\pi j}{n}\right)\right]$$

is relation appears as eqn. (4) in the main text.

In the version of this system with another stage of phase compression, a second-order phase difference $f_2(x)$ is measured given by

$$f_2(x) = \psi_{k-1} - 2\psi_k + \psi_{k+1}$$

By a process similar to the above one, the indicated bearing in this case is found to be

$$\gamma = \arg\left[-\sum_k (\psi_{k-1} - 2\psi_k + \psi_{k+1})\exp\left(\frac{2k\pi j}{n}\right)\right] \quad (19)$$

This can also be simplified to give

$$\gamma = \arg\left[\sum_k \psi_k \exp\left(\frac{2k\pi j}{n}\right)\right]$$

just as in the basic regime of the system.

DISCUSSION ON

"DESIGN OF A LOGARITHMIC RECEIVER"*

Mr. G. Meechan (*communicated*): The design procedure depends on x, the rate of rise which is different for each target return. How then does the author choose x to design his delay line?

The cathode circuit time-constant, when made small compared to a pulsewidth, may lead to negative feedback in the case of short pulses. I have found a value of cathode time-constant at least five times the reciprocal of the i.f. bandwidth necessary to avoid distortion of the trailing edge as shown in Fig. D.

I think it can be shown that one can dispense with the delay line.

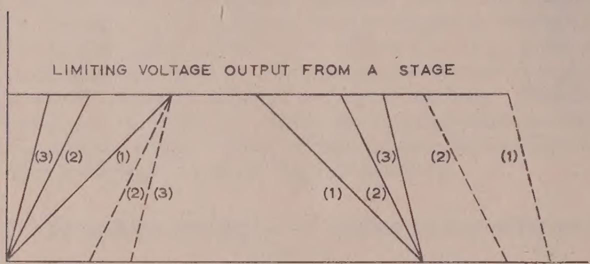


Fig. A

Fig. A shows the output from three stages. The broken curves show the outputs from stages 2 and 3 after having been delayed in order that the leading edges should add up correctly.

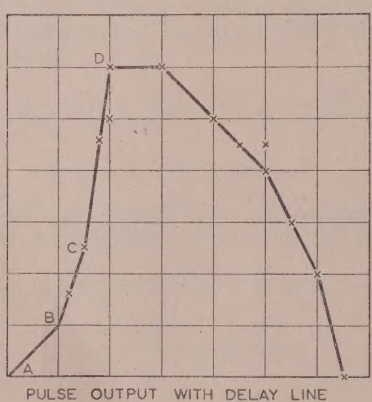


Fig. B

Fig. B shows the shape of the output when the outputs from the stages are added using a delay line.

Fig. C shows the output, when the outputs from each stage are added with no delay line.

It can be seen that portions AB, BC, CD of the leading edge are the same in both cases except in different order. This is to be expected, since a glance at Fig. A will show that we are adding the three leading edges similarly except that they meet at the top in one case and the bottom in the other. The effect, however, on the trailing edge is very different. In Fig. B we see the effect on the trailing edge of the output pulse of the delay line: it is distorted. Fig. C shows a trailing edge which is a mirror image of the leading edge, and the output pulse is symmetrical.

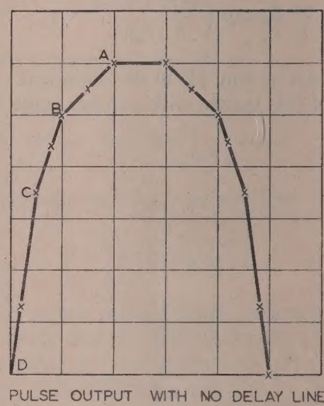


Fig. C

Thus no delay line gives a leading edge as good as that given by a correctly designed delay line: however, no delay line gives a better trailing edge. I therefore submit that there is no ideal design of delay line.

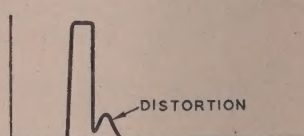


Fig. D

Mr. S. Rozenstein (*in reply*): The slope of the input pulse determines the number of i.f. stages contributing to the final pulse slope. In order to minimize output-slope changes the delay-time must meet some compromise value.

It was found that cathode time-constants larger than pulse-width caused an increase of the recovery time of i.f. stages under overload conditions.

The series delay line may be exchanged for a parallel delay line, in which case the outputs of all the detectors pass through separate delay lines to a common load, but this would complicate the circuit technique and construction. The minimum time-delay in a separate section depends on the amount of i.f. attentuation required to avoid oscillations.

Various leading-edge waveforms can be achieved by adjusting the separate time-delays. The precise waveform is then chosen according to the requirement (automatic tracking, manual tracking, triggering, etc.).

* ROZENSTEIN, S.: Paper No. 1733 R, January, 1955 (see 102 B, p. 69).

PROCEEDINGS OF THE INSTITUTION OF ELECTRICAL ENGINEERS

Part B. RADIO AND ELECTRONIC ENGINEERING (INCLUDING COMMUNICATION ENGINEERING), JANUARY 1956

CONTENTS

Bibliographical References.—It is requested that bibliographical reference to an Institution paper should always include the serial number of the paper and the month and year of publication, which will be found at the top right-hand corner of the first page of the paper. This information should precede the reference to the Volume and Part.

Example.—SMITH, J.: "Reflections from the Ionosphere," *Proceedings I.E.E.*, Paper No. 3001 S, December, 1954 (102 B, p. 1234).

Published by The Institution, Savoy Place, London, W.C.2. *Telephone:* Temple Bar 7676. *Telegrams:* "Voltampere, Phone, London."
Printed by Unwin Brothers Limited, Woking and London.